9-2022

Leveraging Subject Matter Expertise to Optimize Machine Learning Techniques for Air and Space Applications

Philip Y. Cho

LEVERAGING SUBJECT MATTER EXPERTISE TO OPTIMIZE MACHINE LEARNING TECHNIQUES FOR AIR AND SPACE APPLICATIONS

DISSERTATION

Philip Y. Cho, Major, USAF

AFIT-ENC-DS-22-5-002

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT-ENC-DS-22-5-002

LEVERAGING SUBJECT MATTER EXPERTISE TO OPTIMIZE MACHINE LEARNING TECHNIQUES FOR AIR AND SPACE APPLICATIONS

DISSERTATION

Presented to the Faculty

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

in Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy in Applied Mathematics

Philip Y. Cho, B.S., S.M.

Major, USAF

September 2022

AFIT-ENC-DS-22-5-002

LEVERAGING SUBJECT MATTER EXPERTISE TO OPTIMIZE MACHINE LEARNING TECHNIQUES FOR AIR AND SPACE APPLICATIONS

DISSERTATION

Philip Y. Cho, B.S., S.M.
Major, USAF

Committee Membership:

Dr. Aihua W. Wood
Chair

Dr. Andrew J. Geyer
Member

Dr. Brian J. Lunday
Member

Abstract

In this research, we develop machine learning and statistical methods that are tailored for Air Force applications through the incorporation of subject matter expertise. In particular, we develop techniques for incorporating subject matter knowledge in neural networks, Bayesian regression, and structural causal models. These techniques are developed in the context of three separate application areas: localizing point defects in transmission electron microscopy (TEM) of crystalline materials; estimating the relationship between attributes of fighter pilot communities and flight mishap rate; and analyzing Air Force evaluation process.

Our first contribution is a novel method for localizing point defects in TEM images of crystalline materials using principal component analysis (PCA) and a convolutional neural network (CNN). Notably, the design of the PCA-CNN method leverages knowledge about point defects in crystalline materials. Furthermore, the method is a self-supervised method that is trained without labeled images of point defects and, thus, represents a novel methodological contribution. We show that the tailored PCA-CNN method outperforms CutPaste, a state-of-art artificial intelligence (AI) model for defect localization method, on both simulated and experimental TEM images.

Our second contribution reveals the relationship between attributes of fighter pilot communities and flight mishap rates through the use of predictive projection and Bayesian regression. We use personnel and mishap data from 2007-2020 to present an in-depth analysis of historic trends within fighter pilot communities. In our analysis of historic mishap data, we find evidence of abnormal mishap cost estimation behavior near the threshold between class B and C mishaps. Using Bayesian regression with feature selection via predictive projection, we find that pilot communities with higher

average flight hours in the last year are associated with reduced mishap rates. A higher percentages of pilots who are DGs, are IPs, and have advanced academic degrees are also associated with reduced mishap rates.

Lastly, we demonstrate the use of Bayesian priors to incorporate the subject matter knowledge gained from prior qualitative studies on mishap safety. Our third contribution provides a framework for estimating causal effects using data associated with Air Force evaluation processes. Air Force evaluation processes are unique because the causal relationships that induce the observed data are often known. For example, policy dictates which factors can and cannot be considered by a promotion board. We use structural causal models to represent our knowledge of the evaluation processes. Under the assumption of a linear causal model, we derive a formula, $\hat{\beta} = C_x^{-1}\vec{C}'_{xy}$, for computing the coefficients in a regression via the pair-wise covariances of the predictors. This allows for the estimation of causal quantities pertaining to evaluation processes via regression and the proper selection of controls.

To E, J, and E, thank you for your unconditional love and support.

Acknowledgements

The work presented in this dissertation was made possible by the support of numerous individuals. First, I would like to express my deepest gratitude to Prof Aihua Wood for guiding my academic journey and for being a constant source of encouragement. Thank you for your kindness and understanding as I navigated my PhD while managing the many challenges presented by the coronavirus pandemic.

Second, I am eternally grateful to my family for all the sacrifices they have made in support of my PhD. The pandemic created a whirlwind of challenges that were only overcome via the sacrifices made by my wife and kids. Thank you.

To my parents, thank you for all the sacrifices you made to allow me to reach this milestone.

Lastly, I would like to thank the countless individuals who have played a role in my academic journey. There are too many to name, but I would be remiss if I did not acknowledge all the help I have received along the way. Thank you all.

Philip Y. Cho

Table of Contents

Page

List of Figures

List of Tables

LEVERAGING SUBJECT MATTER EXPERTISE TO OPTIMIZE MACHINE LEARNING TECHNIQUES FOR AIR AND SPACE APPLICATIONS

I. Overview

The U.S. Air Force (USAF) Science and Technology Strategy from 2019 states that artificial intelligence (AI) and machine learning (ML) are critical for the development of future strategic Air Force capabilities. Specifically, it states the following,

> To realize the potential of artificial intelligence, the Air Force scientific and technical enterprise must push well beyond developed commercial applications in overcoming major challenges to effective military employment. These include unpredictable and uncertain physical environments, noisy and unstructured data from dissimilar sources, limited training data for machine learning, and the high levels of trust required to support lethal combat operations.

In this research, we seek to"push well beyond" commercial applications of machine learning and statistics by developing methods that are tailored for Air Force applications through the incorporation of subject matter expertise. In particular, we develop techniques for incorporating subject matter expertise in neural networks, Bayesian regression, and structural causal models. These techniques are developed in the context of three separate research thrusts:

- Neural networks for localizing point defects in transmission electron microscopy (TEM) of crystalline materials.
- Bayesian regression for estimating the relationship between attributes of fighter pilot communities and flight mishap rate.

- Structural causal models for analyzing Air Force evaluation process.

In each of these seemingly disparate areas, we show that incorporating subject matter knowledge can enhance the utility of existing machine learning and statistical techniques. Notably, the type of subject matter expertise that is available in each of the application areas is vastly different. In our work with TEM images, we leverage knowledge of the physical phenomena surrounding atomic point defects and crystalline structures to design a neural network model for localizing defects. In our work with fighter mishaps, we show that findings from prior qualitative studies can be incorporated into quantitative models via Bayesian priors. Lastly, we use structural causal models to represent our knowledge of Air Force evaluation processes, which, in turn, allows for the estimation of causal effects. In this chapter, we explain the background, motivation, and key contributions in each of these three research thrusts.

1.1 Thrust 1: Defect Detection in Transmission Electron Microscopy Images via Neural Networks

The first research area involves the use of neural networks to locate atomic defects in crystalline materials, such as semiconductors, using transmission electron microscopy (TEM) images. TEM images reveal the positions of atomic columns in a lattice structure. The properties of crystalline materials are heavily influenced by the presence of point defects. Thus, the engineering of point defects in materials is foundational in the development of novel materials for advanced electronic and photonic applications. A key challenge in the engineering of point defects, however, is determining the location of point defects in the finished crystal. Existing research on the use of machine learning for defect detection in TEM images has focused on supervised learning models where labeled training data is generated using simulated TEM images. Given that simulated images cannot accurately replicate the incon-

sistencies and noise patterns found in experimental images, it is desirable to develop defect detection methods that can be trained on experimental images without labeled examples of point defects. The method we propose is a novel self-supervised convolutional neural network (CNN) model that does not rely on a fully labeled training dataset and is unique in that it can be trained directly on experimental TEM images rather than simulated TEM images. Importantly, the method is designed to take advantage of existing knowledge of the physical phenomena pertaining to crystalline materials. Namely, we leverage knowledge that lattice structures consist of repeating patterns and that defect densities in crystalline materials are relatively low. We show that our defect detection model outperforms a state-of-the-art, general-purpose anomaly detection method on both simulated and experimental TEM images.

1.2 Thrust 2: Influence of Pilot Attributes on Fighter Mishap Rates via Bayesian Analysis

The second research area focuses on the relationship between the attributes of pilot communities and mishap rates. In 2019, in response to numerous high-profile mishaps, Congress commissioned the National Commission on Military Aviation Safety (NCMAS) to assess and identify causes contributing to military aviation mishaps. The NCMAS report, along with multiple other prior qualitative studies, have concluded that flight mishaps can often be attributed to pilot error. However, there has not been a quantitative analysis of the relationship between pilot attributes and mishap rates. Motivated by this shortfall, we first use DOD administrative data to quantify attributes of fighter pilot communities, and then analyze trends within pilot communities from 2008-2020. Given the complexity of the personnel data and numerous data deficiencies that needed to be addressed, our analysis of trends within fighter pilot communities represents a novel contribution. Next, we use fighter mishap data

to provide an analysis of fighter mishap trends. Notably, our analysis of mishap data from 2008-2020 reveals that cost estimates of many class C mishaps are clustered around their upper cost threshold and may have been altered to avoid classification as class B mishaps. As a result, unlike prior studies, we choose to focus our analysis on the combined rate of all class A, B, and C mishaps. We model the association between pilot attributes and annual rate of class A, B, and C flight mishaps, using a Bayesian regression framework. Our results show that prior flight experience, along with several characteristics of a MDS pilot community, are associated with the rate of HCMs. Specifically, we find that MDS pilot communities with 10 more flight hours in the past year are, on average, associated with a 5% lower HCM rate. Additionally, we find that a 0.1 standard deviation increase in the proportion of pilots who are instructor pilots (IPs), distinguished graduates (DGs) from commissioning source, and graduate degree recipients is associated with a reduction in major aviation mishaps by 2.1, 2.0, and 1.3 percent, respectively. In addition to our model results, our use of Bayesian regression and predictive projection for feature selection represents a valuable methodological contribution to aviation accident analysis. Lastly, given the majority of existing research on aviation safety is qualitative in nature, we seek to incorporate the findings of prior qualitative studies in our analysis. We show that Bayesian regression and predictive projection provide an elegant approach for incorporating existing knowledge from prior qualitative studies.

1.3 Thrust 3: Using Causal Inference to Analyze Air Force Personnel Evaluation Processes

The third research area focuses on the use of causal inference theory when analyzing data generated from an Air Force evaluation process. Formal evaluation processes are used throughout the Air Force for a wide range of purposes including selecting

distinguished graduates (DG) from training programs, selecting candidates for professional opportunities, and awards. The results of the evaluation process then affect outcomes of interest such as retention rate, promotion rate, or job performance. The data generated from these processes are commonly used in regression analyses to determine the relationship between various factors and a particular performance outcome. However, in prior studies, there have not been efforts to leverage knowledge of the evaluation process that generated the data used for analyses. Unlike in many other applications, the causal relationships pertaining to Air Force evaluation process are known since they are often dictated by policy. For example, Air Force policy dictates what factors can and cannot be considered for promotion. If we use structural causal models to incorporate knowledge of the evaluation process into our analysis, we show that regression coefficients can be used to estimate causal effects. Notably, we derive and use an alternative formula, $\hat{\beta} = C_x^{-1}\vec{C}'_{xy}$, to compute the regression coefficients of a multiple linear regression using only the pair-wise covariances of the regression predictors. We show that this alternate method for computing regression coefficients allows us to clearly understand which causal quantities are being estimated by the regression coefficients. Therefore, depending on which causal relationship we are trying to estimate, we can determine the set of predictors or covariates that must be controlled for to correctly identify the causal effect of interest.

II. Contribution 1: Defect Detection in Transmission Electron Microscopy Images via Neural Networks

2.1 Overview and Motivation

Crystalline materials, such as semiconductors, are materials comprised of a highly-ordered lattice structure. The properties of crystalline materials are heavily influenced by the presence of point defects in the lattice structure. Thus, the engineering of point defects in materials by the creation of specific defect types and by the control of spatial location and number density is foundational in the development of novel materials for advanced electronic and photonic applications. Transmission electron microscopy (TEM) is a widely used technique for imaging crystalline structures and analyzing point defects due to its versatility for many different modes of imaging and spectroscopy at high spatial resolution. However, detection of point defects in TEM images continues to remain a challenge in many material systems since the contrast due to the defect is affected by various factors such as its local environment and imaging conditions [1, 2, 3]. Figure 1 shows an example of an experimental TEM image. In light of these challenges, the goal of our research is to leverage machine learning methods to accurately locate point defects in crystalline materials using experimental TEM images.

In this chapter, we first propose a novel method for locating point defects in simulated TEM images that uses principal component analysis (PCA) and convolutional neural networks (CNNs). A key advantage of the methodology we propose is that it can be trained using only nondefect TEM images. A defect detection method trained solely using nondefect images is desirable because it is possible to grow experimental crystalline samples that are known to be free of defects. In contrast, when crystalline samples are grown with defects, the true locations of the defects are unknown, so

it is difficult to use experimental defect images to train a model. Thus, a model only trained on nondefect images allows for training via experimental images. Since we are only using nondefect images in the training set, the model is considered a self-supervised anomaly detection model. The model we propose uses PCA to generate a lower-dimensional reconstruction of a TEM image and then a CNN to classify whether a resiudal image contains a defect. We show that, by jointly using PCA and a CNN, we can accurately locate realistic defects in simulated TEM images even in the case where there is significant imaging noise. Our method for locating defects in TEM images has been published in MDPI Mathematics.

Second, we provide an in-depth comparison between the PCA-CNN defect detection method and an alternative, state-of-the-art method for defect detection and localization called the CutPaste model [4]. After training a CutPaste model using TEM images, we show that the PCA-CNN model substantially outperforms the CutPaste model in localizing defects in simulated TEM images. Notably, the design of the PCA-CNN method incorporates existing knowledge about crystalline structures and point defects. In comparison, alternative "off-the-shelf" methods for defect detection do not incorporate any knowledge of about crystalline structure. Lastly, we demonstrate the use of the PCA-CNN model on an experimental TEM image. We show that, despite being trained on simulated TEM images of Gallium-Arsenide (GaAs), the performance of the PCA-CNN model generalizes to an experimental images of a non-GaAs crystalline material.

2.2 Related Work

In recent years, CNNs have proven to be a highly effective tool for image analysis. Applications include image classification, object detection, pose estimation, and text recognition [5]. Given the data-intensive nature of TEM imagery, CNNs have also

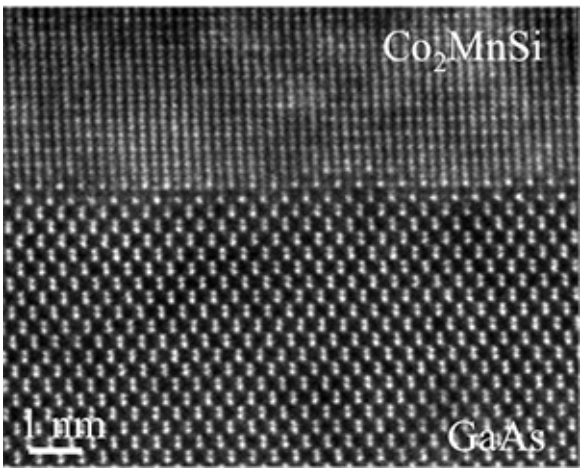


Figure 1. Sample experimental TEM image. In this image, the bright spot corresponds to atomic positions. Note that under different imaging conditions, the color of the atomic positions have be inverted.

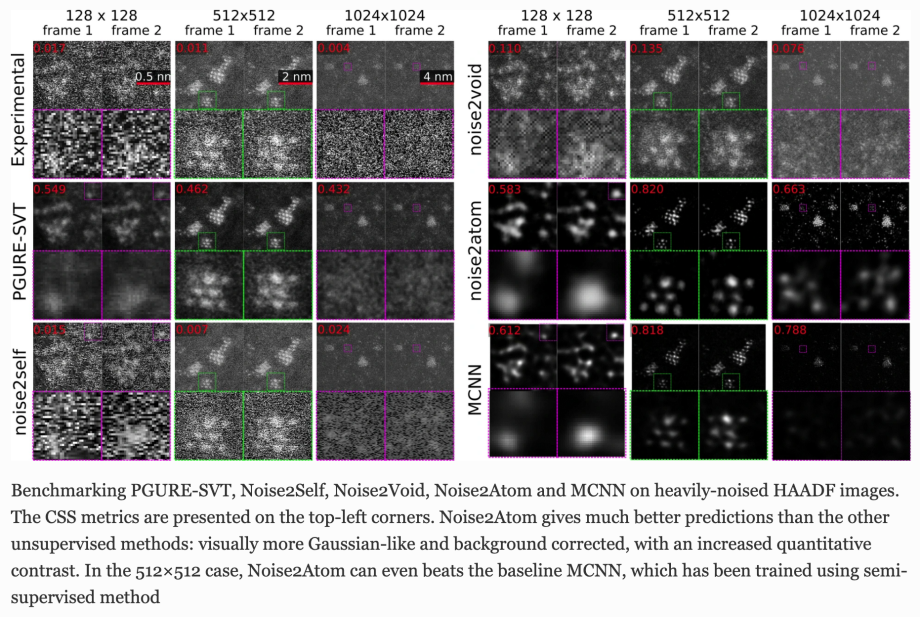
been a useful tool in TEM image analysis. Examples of using neural networks for analyzing TEM images include using CNNs for denoising TEM images [6, 7], generating TEM images from partial scans [8], classifying types of crystalline structures [9], locating defects in non-crystalline materials [10], mapping atomic structures and defects [11], and mapping general structures of interest [12].

We first focus on the latter two studies [11, 12] because they directly address the problem of defect detection in TEM images of crystalline materials using machine learning. In both studies, the framework is to train a CNN using simulated TEM images and then apply the trained models to experimental images. Additionally, both propose training a multi-class classification CNN that outputs pixel-wise classifications. That is, every pixel in a TEM image is assigned a predicted class. In one study [11], the three classes are vacancies, dopants, and defect-free. In another study [12], the classes are general, non-overlapping structural characteristics such as the column height of the sample. Both of these models require extensive simulated data where the true label for each pixel is known. After training the pixel-wise classification model with pixel-by-pixel truth data, the models are shown to produce strong

results on experimental TEM images. Similar to the aforementioned work, we seek to develop a model that can detect local structures of interest in TEM images, namely defects, in crystalline materials. However, Ziatdinov et al. [11] and Madsen [12] both acknowledge the difficulty in acquiring experimental images where the true defect locations are known and accordingly propose models solely trained on simulated data with known defect locations. These models are classification models since they are trained using labeled data. Our focus is on develop anomaly detection methods that do not rely on labeled training data.

The presence of inconsistent imaging noise is a significant differentiator between simulated and experimental images. Thus, research pertaining to the denoising of experimental TEM images is of key interest. Recent work in denoising TEMs via machine learning include using a U-Net CNN to remove Poisson noise from simulated images [7] and using a two-stage generative model trained using both simulated an experimental data [6]. The latter study [6] is particularly interesting because it is, as far as we know, the only denoising model that incorporates experimental data in the training process. Specifically, they propose a method that trains a CNN using clean simulated images and noisy experimental images. Note that simulated noise is not a feature of the training data. The data is used to train a generative adversarial network (GAN) architecture that relies on four submodels. The model ultimately takes an experimental image as an input and outputs a denoised image. The denoising model is referred to as Noise2Atom. The model results are compared to various denoising algorithms and is shown to outperform existing methods. Interestingly, they show that the unsupervised model can, in certain cases, outperform a supervised method. Figure 2 shows the results of the Noise2Atom model in comparison to other methods. Apart of neural network based denoising methods, there have been efforts to use modified PCA methods for TEM denoising [13] and using block-matching to denoise

TEM images [14]. Each of these aforementioned works focus on addressing noise patterns specific to experimental TEM images. Looking beyond TEM images, there are numerous related works on general image denoising [15, 16, 17, 18].



Benchmarking PGURE-SVT, Noise2Self, Noise2Void, Noise2Atom and MCNN on heavily-noised HAADF images. The CSS metrics are presented on the top-left corners. Noise2Atom gives much better predictions than the other unsupervised methods: visually more Gaussian-like and background corrected, with an increased quantitative contrast. In the 512×512 case, Noise2Atom can even beats the baseline MCNN, which has been trained using semi-supervised method

Figure 2. See caption above from Wang, et. al. [6]

Lastly, we review other related works. In addition to containing imaging noise, experimental TEM images can often suffer from poor contrast conditions where it is difficult to distinguish atomic structures. Thus, several works propose deep learning models for enhancing the resolution of scanning electron microscopy (SEM) images [19, 20]. A resolution enhancement model may be helpful as a pre-processing step when trying to locate defects in TEM images since it can correct any local distortions due to imaging conditions. Neural networks have also used to generate full-scale TEM images from an incomplete, partial STEM image [8]. A generative neural network can be used to "fill-in" areas that were not imaged. Models that complete missing portions of an image have also been shown to be useful for anomaly detection [21] in non-TEM applications.

The final related work we review is a recent paper from Google that addresses defect detection and localization in a broad range of images "CutPaste" [4]. A self-supervised CNN model is trained to classify "normal" and defect images through the use of randomly inserted "CutPaste" defects. This paper is of particular interest because, although it is not specific to TEM images, the methodology they propose has much in common with the methodology we have used in our analysis of TEM images. The CutPaste algorithm is also self-supervised since it only needs training data that is free of defect. While there are other existing defect detection methods, we focus on the CutPaste model because it outperforms existing anomaly detection methods on open source benchmark datasets.

The related works presented in this section demonstrate the current state-of-the-art applications of machine learning with TEM images.

2.3 Defect Detection in Simulated TEM Images

In this section, we introduce methods for localizing point defects in simulated TEM images. We first present the data and two separate methods for defect detection. We first propose a method for locating defects in a TEM image using PCA-based reconstruction error. We show that the PCA-based defect detection model performs well in the case of no imaging noise, but performance deteriorates in the presence of imaging noise. We improve upon the PCA-based defect defection model with a weakly supervised CNN and show that the combined approach, referred to as the PCA-CNN method, improves defect localization performance, particularly in the case of high levels of imaging noise.

2.3.1 Data

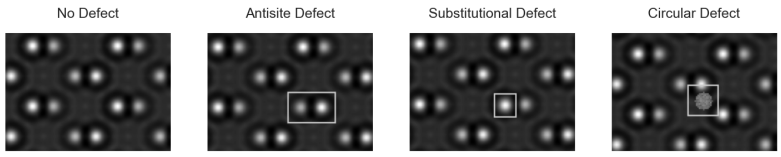
The first step in developing a model for predicting the location of point defects is to generate simulated TEM images. TEM images for GaAs were simulated using the TempasTM software for a crystal projected along the (110) zone axis for TEM accelerating voltage of 300kV and up to specimen thickness of 15nm. The imaging parameters for the objective lens were set such that the spherical aberration coefficient was $-15\,\mu$m and defocus ranging from -20nm to $+20$nm.

Ideally, experimental data would be used for this study, but due to the difficulty in acquiring experimental data, we utilize simulated TEM images to train and test our defect detection models. The use of simulated data is a start towards developing a method that can be trained directly on experimental data. A key consideration, then, is an understanding of the extent to which we can control defects in experimental images. As discussed earlier, it is possible to produce experimental GaAs samples that are defect-free so we assume it is feasible to acquire experimental TEM images that are known to be defect-free. In contrast, when defects such as dopants are added to experimental GaAs samples during the production process, the true locations of the dopant atoms in the GaAs sample are unknown. Thus, it is infeasible to generate a set of TEM images for which we know the true location of the point defects. The lack of knowledge about the true location of the defects in an experimental image is crucial. In light of this lack of defect truth data, the goal is to develop a defect detection method trained solely on defect-free TEM images. Our dataset consists of simulated TEM images of GaAs using 8 different thickness conditions and 21 different defocus conditions. The thickness is varied from 1nm to 15nm in 2nm steps. The defocus condition ranges from -20nm to 20nm in 2nm steps. Thus, there are a total of 168 unique imaging conditions. These 168 imaging conditions are split into a set of 112 train conditions (66%) and 56 test conditions (33%). The splitting of the train and

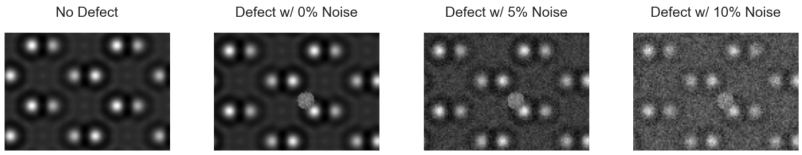
test conditions is done in a nonrandom manner. A third of the defocus conditions, {−18nm, −12nm, −6nm, 0nm, +6nm, +12nm,+18nm}, are assigned to the test set and the remaining conditions are assigned to the training set. The imaging conditions have a significant impact on the resulting TEM image (Figure 6), so splitting on the imaging conditions ensures that model performance generalizes beyond conditions only in set of training conditions. For the remainder of the paper, we refer to these sets as the train and test conditions.

We use the train and test conditions to further generate the training and tests data for our models. For each of the 112 train conditions, we simulate a single TEM image of dimension 1007×1024. The image is represented as a matrix of dimension 1007×1024 where each entry represents a grayscale pixel value. Since the TEM image consists of a repeating lattice structure, we choose to analyze the TEM images in smaller segments of dimension 84×118. Each of these image segments is large enough to include two sets of GaAs pairs in both the vertical and horizontal direction. At the same time, these image segments are small enough such that accurately identifying the presence of a defect in a particular image segment is nearly equivalent to determining the location of the defect. Thus, after generating the larger simulated TEM images, we generate 50 random crops from each training set image where each crop is an image segment of dimension 84×118. Note that the crops are random, so the location of the GaAs atoms differs within each image segment. These 5600 image segments constitute the training data for the PCA and form the basis for the training data for the CNN.

Next we use the test conditions to generate the test data. For each of the 56 test conditions, we generate 30 TEM images that are each 1007×1024. Specifically, each simulated image contains a single point defect that can be one of three defect types. For each of these three defect types, 10 replicates are generated wherein the defect location is randomized for each replicate. This results in a total of 30 simulated



(a) Examples of defect types



(b) Examples of effect of increasing imaging noise

Figure 3. (3a) Three different types of defects are considered. For each imaging condition in the test set, each of the three defect types is added to the test image. (3b) Examples of increasing levels of Gaussian noise. The noise percentage level corresponds to the variance, σ^2, of the Gaussian noise that is added the image. A circular defect is shown for reference.

images for each test condition. The three types of defects are 1) an antisite complex where the Gallium and Arsenic atoms are reversed, 2) substitutional defect where a dopant has an approximately 5% larger radius, 3) an arbitrary circular defect. Figure 3a shows an example of each of the three defect types. We choose to consider these three types of defects because it includes a very subtle defect in the substitutional defect, a more obvious defect in the antisite defect, and a general defect in the circular defect. The circular defect is located randomly in an image segment while the other two located appropriately. The circular defect represents any general point defect such as an interstitial defect or a vacancy. The circular defect is unique in that it is easily added to any TEM image, either simulated or experimental. This flexibility plays an important role in the CNN model that is introduced in a later section. For each combination of imaging condition and defect type, we generate 10 simulated TEM images with a randomly located defect. This results in 1680 test images where the defect location is known. Unlike the smaller image segments used in the training set, the images in the test set are 1007×1024. The test set images are used to evaluate

whether or not the defect detection methods can accurately predict the location of the defect in the test image. The process for generating the training and test data is summarized in Figure 4.

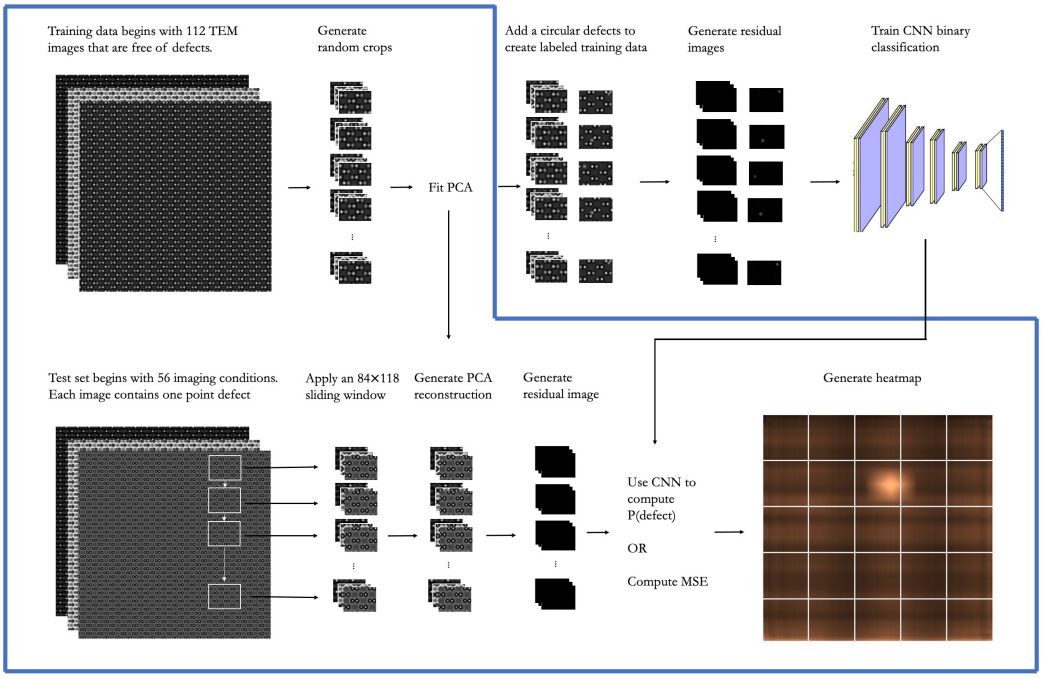


Figure 4. The methodology of the PCA based defect detection method is summarized by the steps bordered in blue. The steps outside the blue border are the additional necessary to incorporate the CNN classification model into the defect detection methodology.

The simulated TEM images do not include imaging noise. However, experimental TEM images can have varying degrees of noise that make it difficult to identify defects in a TEM image. Therefore, it is desirable for our proposed defect detection methods to be robust to imaging noise. To account for the presence of imaging noise in experimental images, Gaussian noise is used in both the training and test sets. Specifically, Gaussian noise with $\varepsilon \sim \mathcal{N}(\mu = 0, \sigma^2 = 0.05)$ is added to each pixel value for images in the training set. For the test set, varying levels of Gaussian noise, where $\sigma^2 = 0.00, 0.05, 0.10$, are added to the TEM images and model performance is evaluated for each noise level. Figure 3b shows the effect of the Gaussian noise on a

TEM image.

2.3.2 PCA Model

We present a method of detecting defects using PCA reconstructions. We fit a PCA transformation on the 5600 defect-free 84×118 image segments in the training set. Then we apply an 84×118 sliding window across each 1007×1024 test set image and, for each window, we generate a PCA reconstruction of the image segment in the window. Since the PCA transformation (and inverse transformation) is only fitted on defect-free TEM images, the assumption is that PCA will struggle to reconstruct an image with a defect. Thus, we expect that the reconstruction error of image segments with a defect to be greater than the reconstruction error of images without defects. We can predict the location of a defect by identifying the image segment with the highest MSE. With this general framework in mind, we present the method in more detail below.

For our PCA-based model, the training data consists of 50 randomly cropped image segments from the each of the 112 larger TEM images in the training set. These 5600 training image segments can be represented by the matrix $\mathbf{Q} \in \mathbb{R}^{5600 \times 9912}$ where the rows represent individual image segments and the columns represent mean-centered values at each pixel location. The orthogonal linear transformation $\mathbf{Q_k} = \mathbf{QW_k}$ projects the original data, $\mathbf{Q}$, to a lower k dimensional representation, $\mathbf{Q_k}$. In PCA, the weight matrix $\mathbf{W_k} \in \mathbb{R}^{9912 \times k}$ is constructed such that the reconstruction MSE, $\left\|\mathbf{Q} - \mathbf{Q_k}{\mathbf{W_k}}^T\right\|_F^2$, is minimized. Notice that $\hat{\mathbf{Q}} = \mathbf{Q_k}\mathbf{W_k^T}$, a matrix of dimension 5600×9912 represents the reconstructed images. The projection to the lower dimensional space and the reconstruction back to the original dimensional space are both determined by $\mathbf{W_k}$. Once $\mathbf{W_k}$ is fit using the training data, it can be used to generate the reconstruction of any 84×118 image segment.

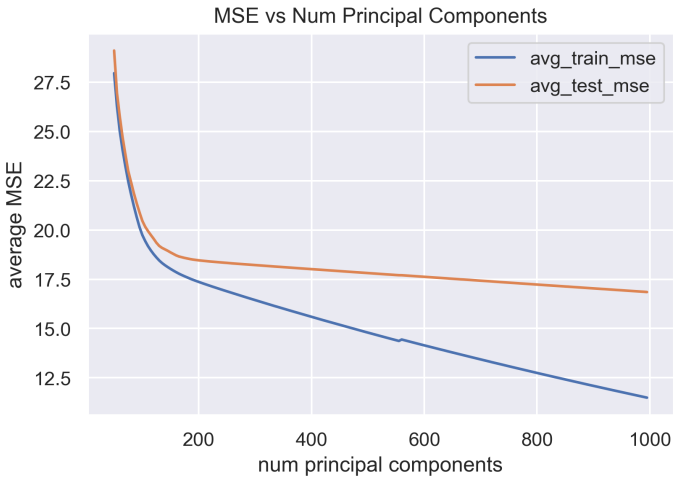


Figure 5. The number of components used to fit the PCA is determined using the average reconstruction MSE of the test set images. The average reconstruction MSE for test set images falls rapidly and levels off after the number of components exceeds 150.

We set the value of k using reconstruction mean-squared error (MSE). Specifically, we fit the PCA using the 5600 image segments in the training set and then apply the fitted PCA to image segment from the test conditions to compute the average reconstruction MSE. For each of the 56 test conditions, 50 random crops are taken where each crop is known to be free of defects. Figure 5 shows the effect of increasing the number of components on MSE. To prevent overfitting to the noise in the training set, we set $k = 150$. Figure 5 shows several examples of a image segments under various imaging conditions as well as the associated reconstruction with $k = 150$. Figure 6 also shows examples of circular defects and effect of the PCA reconstruction on the defect. The circular defects in the raw image are not visible in the PCA reconstruction which indicates that PCA reconstruction struggles to accurately reconstruct anomalous point defects.

The difference between an image segment and its reconstruction is referred to as the residual image. The residual image, intuitively, shows what is remaining when the general lattice structure is "subtracted" from the original image. Thus, the residual

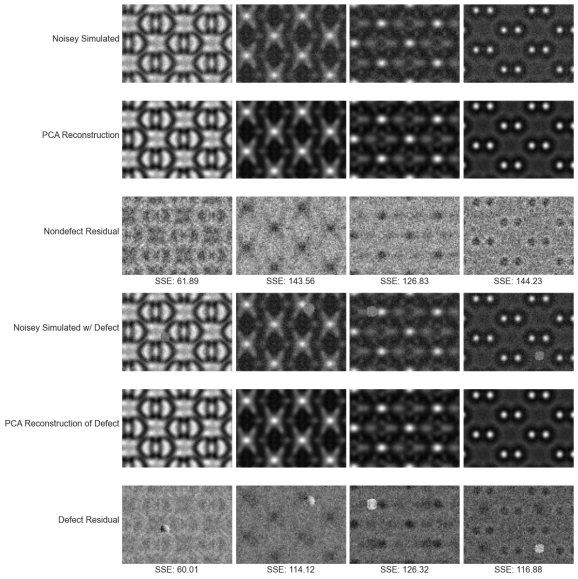
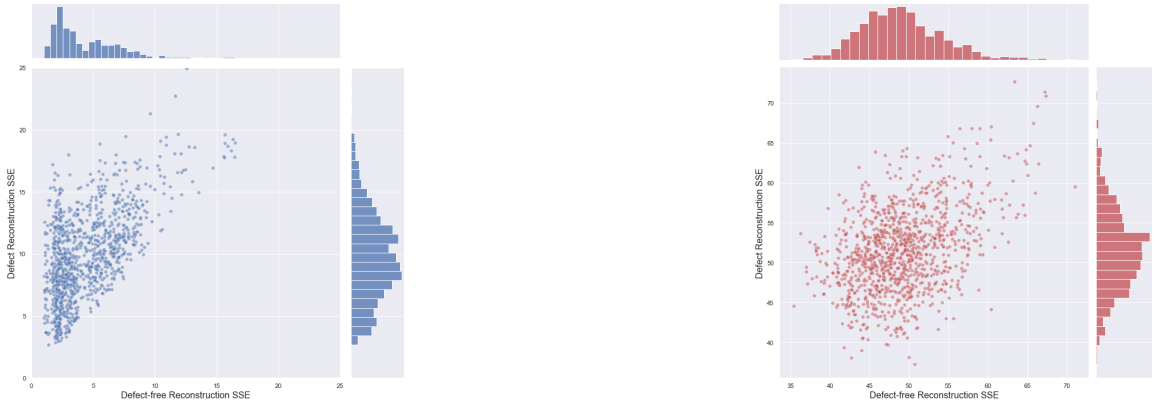


Figure 6. The first three rows show 1) the defect-free image segments with noise, 2) the PCA reconstruction, and 3) the residual between the raw and reconstructed image, respectively, for a range of imaging conditions. The bottom three rows show the same sequence images except the raw image contains a circular defect that has been randomly inserted. Notably, the PCA reconstructed image does not accurately reconstruct the defect since the PCA transformation was fitted only on images without defects.

images consists of noise and any anomalies in the lattice structure. The reconstruction MSE can be regarded as a scalar that summarizes the residual image. For each of the 5600 images in the training set, we can compute the reconstruction MSE with and without a circular defect to understand the distribution of reconstruction MSE. Figure 7a shows how the presence of a defect changes the reconstruction MSE for each training example. In addition, Figure 7b shows how the addition of imaging noise affects the reconstruction MSE distribution with and without a defect. The concept of a residual image plays an important role in the CNN model that is presented in the next section.

After fitting the PCA transformation, we apply the resulting $\mathbf{W_k}$ to the test set images via a sliding window. Recall that each test set image is of dimension 1007×1024 and contains a single point defect with known location. We use a 84×118 sliding window across the 1007×1024 image and, for each window, we complete the following



(a) without noise **(b) with noise $\sigma^2 = 0.10$**

Figure 7. The scatterplot shows the reconstruction MSE of 5600 defect-free image segments (x-axis) in the PCA training set and the corresponding reconstruction MSE for the same image segment with a circular defect inserted. Points that are close to the $x = y$ represent image segments where the reconstruction MSE does not differ much with or without a defect. The marginal plots show the distribution of reconstruction MSEs with and without defects. On left, without imaging noise. On right, with imaging noise.

three steps: 1) generate the PCA reconstruction, 2) generate the residual between the original image segment and the reconstruction, 3) compute the pixel-wise mean squared error (MSE). We then generate a heatmap that shows the average reconstruction MSE for each pixel in the full-size TEM image. The predicted location of the defect corresponds to the area of the heatmap that has the largest reconstruction MSE. Figure 8 shows an example of a test image and the corresponding MSE heatmap. The defect in the test image is a substitutional defect where a single Gallium atom is replaced with a dopant atom that has a 5% larger radius. The defect is difficult to identify visually, but the heatmap accurately locates the defect. This method is applied to all imaging conditions in the test set and we evaluate the accuracy in predicting the location of each type of defect. Figure 4 summarizes the process for predicting defect location using PCA.

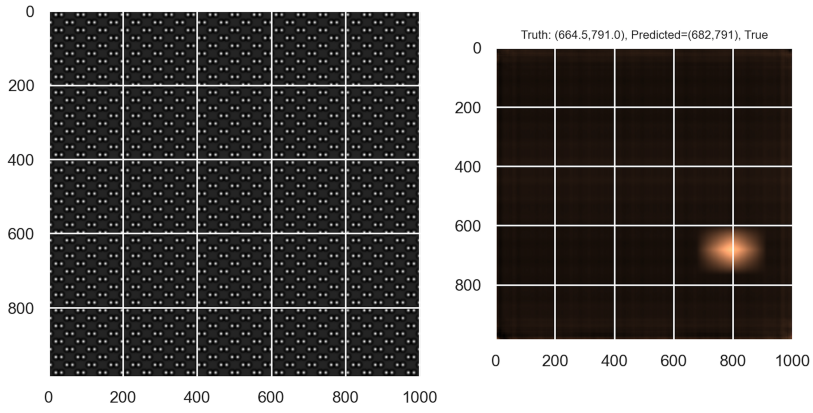


Figure 8. Heatmap that shows the pixels with the largest average MSE based on the PCA reconstruction. Bright spots correspond to areas that are mostly likely to have a defect.

2.3.3 PCA-CNN Model

In this section, we supplement the PCA-based detection method with a CNN classifier to improve the accuracy of the defect location predictions. This combined method significantly improves the prediction accuracy of the PCA model, especially in the case when there is imaging noise.

The PCA-based defect detection method has the benefit of being straightforward. However, in the presence of imaging noise, using PCA reconstruction error can lead to issues. Figure 6 shows the PCA residual images of segments with and without defects. In these particular examples, the reconstruction MSE for the defect images is actually lower than the reconstruction of the MSE for the defect-free images. Notably, if we visually inspect the residual images, the residual images clearly show the presence of a point defect. To address this shortcoming, we introduce a CNN classification model fitted on the PCA residual images. Intuitively, reconstruction MSE is equivalent to adding up the squared values in the residual image, and it ignores any local patterns in the residual image. Alternatively, A CNN can be trained to look for the presence of local patterns in the residual image that may be evidence of a defect. To the best of

our knowledge, the use of the residual image for defect detection is a novel approach.

The training data for the CNN model begins with the same set of defect-free training images used to fit the PCA. Recall that 50 random crops from each of the 112 training images were used to fit the PCA. These same 5600 images are used to build a set of labeled training data for the CNN classifier. Since the training data only includes image segments that are defect-free, a set of labeled training data with defects is generated by adding random, circular defects to each of the 5600 training images. The use of circular defects is motivated by the knowledge that point defects are generally circular in nature. These synthetic defects could be representative of an interstitial defect or a vacancy, but they are not necessarily meant to represent a realistic defect that would be observed in an experimental image. Instead, the objective is for the CNN to classify any residual image with an circular abnormal local pattern as one containing a defect. Since the circular defects are arbitrary and are added post-hoc to the simulated image, this method can easily be applied to experimental TEM images as well. After generating the labeled training, a CNN classification is trained such for an input PCA residual image, the model outputs a scalar $\hat{y} = \mathrm{P}(\text{defect})$ where $\mathrm{P}(\text{defect}) \in [0, 1]$ is the probability that the image segment contains a defect. A summary of the CNN model development process is depicted in Figure 4.

Our CNN architecture is adapted from the classic LeNet-5 architecture . Fig 9 shows the details of each layer of the CNN. It contains four convolutional layers with max-pooling following by two dense layers. We use a binary cross-entropy loss function and is optimized using nAdam. The model is fitted for 400 epochs. Importantly, the training data is generated randomly for each batch so the location of the circular defects in the training set are randomized during training. The CNN is trained using Python 3.7 and Keras 2.3 with a TensorFlow 2.4.1 backend. The model achieves

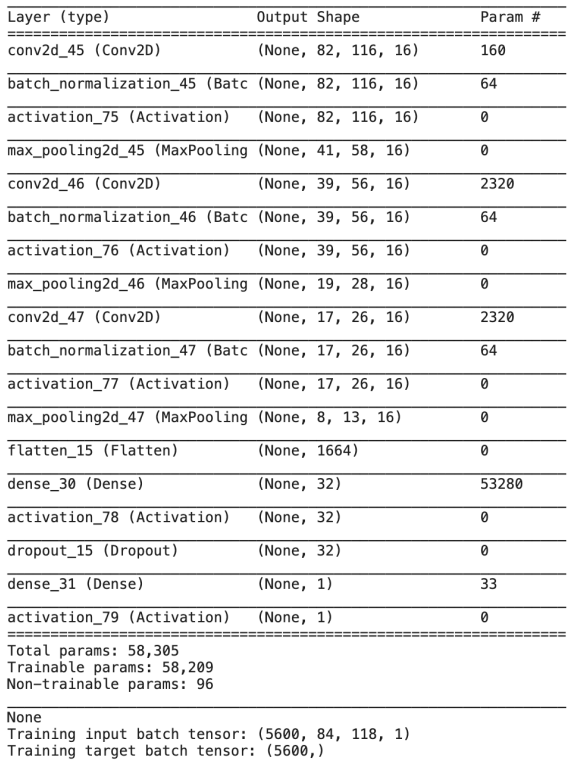


Figure 9. CNN architecture used in the PCA-CNN method. The architecture is motivated by LeNet-5.

$> 99\%$ training and validation accuracy in less than 100 epochs.

After training the CNN, an 84×118 sliding window is applied to each test image. For each 84×118 window, we apply the following three steps: 1) generate a PCA reconstruction, 2) generate a residual image between the original image segment and the PCA reconstruction, and 3) pass the residual image into the trained CNN to generate P(defect). For each pixel in the 1007×1024 test image, we compute the average P(defect) for all sliding windows that contain the pixel. This results in a smoothed heatmap for the entire test image. The location of the defect is then predicted to be the area of the heatmap that has the highest average P(defect). The heatmap shown earlier in Figure 4 is an example of a heatmap generated using the CNN classification model with a sliding window.

In many applications of CNNs for anomaly detection, the output of the CNN

classifier, P(defect) , is compared to a fixed threshold value to determine if a particular input contains an anomaly or not. Notice that a threshold is not necessary here since the predicted defect location is simply the pixel value with the largest average P(defect) . If we generalize to the case where there are n defects in a GaAs sample, then the locations corresponding to the n largest average P(defect) would be the predicted locations of the defects.

2.3.4 Results

We compare the performance of the two defect detection methods discussed above. Recall that there are 56 imaging conditions that were reserved for the test set and there are three defect types. For each combination of imaging condition and defect type, we generate 10 simulated TEM images, each of dimension 1007×1024, where the defect location is randomized. This results in 1680 test images where the defect location is known. For each of the 1680 test images (540 images for each of the three defect types), we apply the PCA and PCA-CNN defect detection methods to predict the location of the defect. We compare the predicted defect location to the true defect location to determine whether the model successfully located the defect.

Table 1a shows the accuracy of both methods in predicting the defect location for various levels of imaging noise. The PCA defect detection method performs particularly well in the case of no imaging. It accurately locates all three defects types at nearly $> 97\%$ and generally outperforms the CNN model. However, as the imaging noise increases, we observe the superior performance of the CNN model. Specifically, when imaging noise rises to $\sigma^2 = 0.10$, the PCA model achieves an accuracy of 56% and 57% on antisite and circular defects, respectively, while the CNN model achieves 75% and 93% accuracy.

The results in Table 1a report the performance of the two methods under all

Table 1. Accuracy of the PCA and PCA-CNN model in locating point defects in the test set images. Table 1a shows the accuracy results when including all images in the test set. Table 1b shows the accuracy results when only the nominal defocus conditions are included. In both cases, the CNN model is more robust to imaging noise.

(a) Location detection accuracy including all imaging conditions.

Method	Noise	Substitution $n = 540$	Antisite $n = 540$	Circular $n = 540$
PCA	σ^2=0.00	0.97	1.00	1.00
	σ^2=0.05	0.16	0.80	0.94
	σ^2=0.10	0.04	0.56	0.57
PCA-CNN	σ^2=0.00	0.71	0.86	1.00
	σ^2=0.05	0.64	0.90	0.99
	σ^2=0.10	0.14	0.75	0.93

(b) Location detection accuracy for central defocus conditions, { -6nm, 0nm, +6nm }.

Method	Noise	Substitution $n = 240$	Antisite $n = 240$	Circular $n = 240$
PCA	σ^2=0.00	1.00	1.00	1.00
	σ^2=0.05	0.24	0.91	0.92
	σ^2=0.10	0.04	0.70	0.61
PCA-CNN	σ^2=0.00	1.00	0.98	1.00
	σ^2=0.05	0.89	0.99	0.99
	σ^2=0.10	0.23	0.89	0.91

test imaging conditions. Recall that the test set includes an equal number of TEM images for a range of defocus conditions. In practice, extreme defocus conditions are relatively uncommon and are actively avoided. Narrowing the focus to the central range of defocus conditions, { -6nm, 0nm, +6nm }, provides a better representation of expected performance on experimental images. Table 1b shows the defect location accuracy of both methods under nominal defocus conditions. Under the restricted set of defocus conditions, the CNN model remains more robust in the presence of imaging noise. Specifically, when $\sigma^2 = 0.10$, the CNN model achieves 89% and 91% accuracy for antisite and circular defects, respectively, while the PCA model achieves 70% and 61% accuracy.

Based on these preliminary results, it appears that the substitution defects are more challenging to identify compared to the antisite and circular defect. This result

is unsurprising given substitution defects are also the most challenging to identify from visual inspection. The substitution defects were purposely subtle so as to determine the effectiveness of the proposed methods for a wide range of defects. In practice, the substitution defects are unlikely to sit precisely in a gallium or arsenic site. If the substitution defect is slightly misaligned, then it is likely that the proposed methods would be more effective in locating the defect. The antisite and random circular defects are more readily identified visually which is reflected in the accuracy results. Although the circular defect is not representative of a particular defect, the circular defect could be representative of an interstitial defect or a vacancy.

2.3.5 Discussion

We introduce two methods for determining the location of a point defect in an TEM image of GaAs. Compared to recent applications of using CNNs for defect detection (e.g., [11], [12], and references therein), the proposed PCA and PCA-CNN methods of defect detection are unique in that they can be trained on TEM images that are defect-free. Unlike prior approaches to defect detection, this opens the door to training these models using experimental data. After training both models using a set of simulated images that are free of defects, we demonstrate the performance of both methods in locating a simulated defect in an HRTEM image. In the case of no imaging noise, we show the PCA method is sensitive to minor defects such as a subtle substitution defect (97% accuracy). However, as imaging noise is introduced, the performance of the PCA method declines rapidly. Supplementing the PCA method with a CNN classification model improves the performance of the model dramatically. The CNN classification model achieves $> 89\%$ accuracy for both antisite and circular defects at the highest level of imaging noise ($\sigma^2 = 0.10$). These results suggest that the CNN approach has the potential to be highly effective in analyzing experimental

images.

Our PCA-CNN classification model is unique in that it is trained on PCA residual images. Using the PCA reconstruction to generate a residual image is a novel approach that has notable benefits. One of the benefits is that it allows for a single pre-trained CNN to be used for a wide range of crystalline materials and imaging conditions. This is in contrast to prior studies that require models trained for specific materials and conditions. Imaging conditions, such as thickness and defocus condition, change the overall "pattern" that is visible in an TEM image. By taking the difference between an image segment and its reconstruction, we are, intuitively, "subtracting" the pattern that is associated with a set of imaging conditions. The residual images are then uncorrelated with the imaging conditions used to generate the TEM image and can be analyzed using a single pre-trained CNN. Another benefit is that using the residual images allows a CNN to more effectively classify defects. Specifically, when we trained a CNN classification model directly on image segments in the training set without using residual images, the trained model far under performed our model that uses residual images. This suggests that use of residual images is a key step in training an effective CNN classification model in the context of TEM images.

The work discussed thus far in this chapter is published in *MDPI Mathematics* with the following citation:

> C.; Wood, A.; Mahalingam, K.; Eyink, K. Defect Detection in Atomic Resolution Transmission Electron Microscopy Images Using Machine Learning. Mathematics 2021, 9, 1209. https://doi.org/10.3390/math9111209.

In the following sections, we expand on this work by comparing the PCA-CNN model to a state-of-the-art, general-purpose anomaly detection method named CutPaste. We also apply the PCA-CNN method to an experimental TEM image.

2.4 Comparing PCA-CNN with CutPaste

The PCA-CNN method outlined in Section 2.3 was designed specifically for localizing point defect in TEM images of crystalline materials. Importantly, the PCA-CNN was designed to take advantage of characteristics of crystalline materials, TEM images, and point defects. It is not designed to be a general purpose defect detection method. In this section, we compare the PCA-CNN method with a general purpose defect localization method called CutPaste [4]. We focus on the CutPaste method, which was published by researchers at Google Cloud AI in 2021, because it is a recently developed method that has achieved state-of-the-art performance on the MVTec benchmark dataset for anomaly detection. In general, we show that PCA-CNN method outperforms the CutPaste method in localizing point defects in TEM images both with and without imaging noise.

We first train a CutPaste CNN model using simulated TEM images known to be free of defects. Second, we evaluate the ability of the CutPaste model to localize simulated defects on the same set of test images described in Section 2.3. Specifically, we evaluate whether the trained CutPaste model can localize simulated antisite, substitutional, and interstitial point defects. Lastly, we discuss the differences between the two methods and consider the manner in which the PCA-CNN benefits from leveraging knowledge about TEM images and point defects. Although general purpose methods like CutPaste may perform well on benchmark datasets like MVTec, we emphasize the incorporating knowledge of the particular phenomena or application under consideration.

2.4.1 CutPaste Model

The CutPaste method is a general purpose defect detection and localization method that is trained only using normal, or defect-free, training data. The method relies on

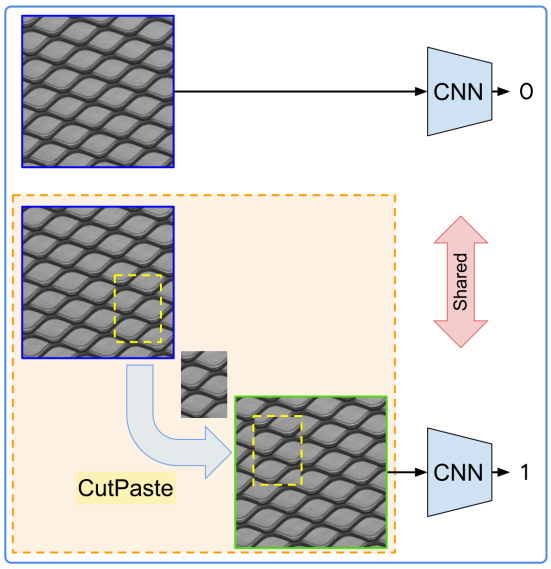


Figure 10. The CutPaste model is trained using a set simulated TEM images that are free of defect. A random rectangular section of the defect-free image is copied and pasted to generate a pseudo-defect. A CNN is then trained to classify images as either having or not having a defect

a novel data augmentation technique to create a labeled set of defect samples that are then used to train a classification CNN. The method is considered a one-class classification model since the objective is to classify a sample as either having a defect or not. Similarly, the PCA-CNN method from section 2.3 is also a self-supervised, one-classification model and, thus, warrants comparison the state-of-the-art CutPaste model. The primary difference between the methods is that the PCA-CNN method uses a circular defect for data augmentation and has an additional step that uses a PCA reconstruction to generate a residual image. We first describe the CutPaste algorithm in detail and then train the CutPaste model using simulated TEM images of GaAs.

Training and using a CutPaste model involves the following steps:

1. Generate a set of $N/2$ images that are normal or free of defects. These images are labeled as normal.

2. For each image, copy a rectangular patch and paste it randomly into the image. These images are labeled as having a defect.

3. Train a ResNet-18 CNN with cross-entropy loss to classify images as normal or defective.

4. For a new image with an unknown number of defects, apply a sliding window and input each image into the trained CNN.

5. Map the output of the CNN, $f(x)$, to an anomaly score using a Gaussian density estimator (GDE),

$$\log p_{\text{gde}}(x) \propto \left\{ -\frac{1}{2}(f(x)-\mu)^{\top}\Sigma^{-1}(f(x)-\mu) \right\}$$

   where μ and Σ are estimated using the normal training data.

6. Compare the anomaly score to a threshold to predict whether a window contains a defect.

We train the CutPaste model using the same initial training data described in Section 2.3. However, rather than using randomly located circular defects, we generate a set of defect images by using rectangular patches that are copied and then pasted randomly back into the original image. The set of normal and defect images are then used to train a CNN using cross entropy loss. Whereas the original CutPaste method uses a ResNet-18 CNN architecture, we choose to use a simpler, LeNet-5 type architecture shown in Figure 11. The CutPaste model was originally trained on the MVTec benchmark dataset which is significantly larger and more diverse than our TEM dataset. Figure 12 provides several sample images from the MVTec dataset [22]. Since the TEM dataset is more homogenous than the MVTec dataset, we determined that a ResNet-18, with nearly 100 million trainable parameters, was unnecessary. The

simpler CNN has 113,000 trainable parameters and is trained for 1000 epochs using randomly generated CutPaste defects and randomly added Gaussian noise. After 1000 epoch, the CNN model achieves 96.5% and 96.0% accuracy on the training and validation set. The strong in-sample and out-of-sample performance suggests that the larger ResNet-18 is likely to lend minimal performance gains.

Layer (type)	Output Shape	Param #
conv2d_48 (Conv2D)	(None, 82, 116, 8)	80
batch_normalization_48 (Batc	(None, 82, 116, 8)	32
activation_80 (Activation)	(None, 82, 116, 8)	0
max_pooling2d_48 (MaxPooling	(None, 41, 58, 8)	0
conv2d_49 (Conv2D)	(None, 39, 56, 16)	1168
batch_normalization_49 (Batc	(None, 39, 56, 16)	64
activation_81 (Activation)	(None, 39, 56, 16)	0
max_pooling2d_49 (MaxPooling	(None, 19, 28, 16)	0
conv2d_50 (Conv2D)	(None, 17, 26, 32)	4640
batch_normalization_50 (Batc	(None, 17, 26, 32)	128
activation_82 (Activation)	(None, 17, 26, 32)	0
max_pooling2d_50 (MaxPooling	(None, 8, 13, 32)	0
flatten_16 (Flatten)	(None, 3328)	0
dense_32 (Dense)	(None, 32)	106528
activation_83 (Activation)	(None, 32)	0
dense_33 (Dense)	(None, 1)	33
activation_84 (Activation)	(None, 1)	0

Total params: 112,673
Trainable params: 112,561
Non-trainable params: 112

Figure 11. The original CutPaste model utilizes a ResNet-18. In this work, a simpler CNN is used since it is trained exclusively on TEM images.

In addition to preferring a simpler CNN over the ResNet-18, we further deviate from the original CutPaste model by electing not to use an anomaly score with a threshold. The anomaly score maps the output of the CNN model to the log-likelihood of a Gaussian distribution with parameters that are estimated from the normal training data. Since the parameters are estimated from the training data, using a GDE anomaly score provides a scaled metric that describes how unusual an image is relative to the normal training data. With the MVTec dataset, an anomaly

score is crucial since it is unknown whether an image contains a defect or not, and a threshold must be used to determine whether the model predicts a defect. However, in our TEM application, the defect density within a crystalline material is known in advance since the material properties are carefully controlled during the growing process. If it is known that a material contains defects, then we are only concerned about localizing the defects and do not have to consider the possibility that there are no defects in the sample. Thus, we can use the raw output of the CNN and simply predict the defect locations to be those where the CNN output is largest. Figure 13 provides an example of a TEM image with a CutPaste defect and the corresponding heatmap that is generated using the output of the CNN. As expected, the CutPaste model accurately locates the defect since the defect is precisely the type of defect used in the training set. In the following section, we compare the performance of the PCA-CNN with that of the CutPaste model for more realistic defects that are not included in the training set.

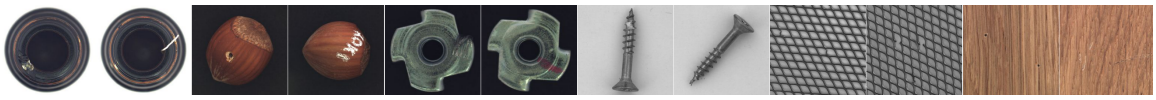
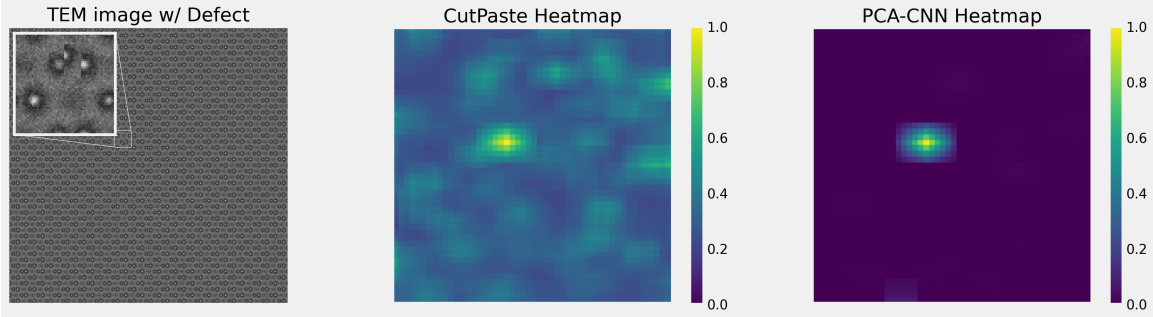


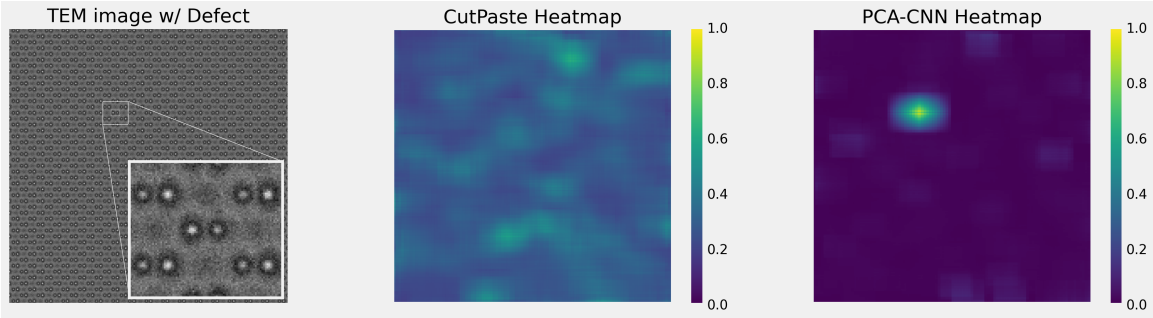
Figure 12. The MVTec dataset consists of 5 different textures and 10 different objects. For each the 15 categories, there are labeled examples of normal and defect images. The CutPaste model achieves state-of-the-art performance on the MVTec dataset.

2.4.2 PCA-CNN vs CutPaste Results

The CutPaste and PCA-CNN models both involve generating training data using data augmentation. The CutPaste model uses random rectangular defects and the PCA-CNN model uses random circular defects to generate set of data that is labeled as having a defect. To compare the performance of both models in localizing more realistic defects, we use the same simulated antisite and substitutional defects that are shown in Figure 3a. As was described in Section 2.3, we add a single simulated



(a) CutPaste defect with noise $\sigma^2 = 0.05$



(b) Antisite point defect with noise $\sigma^2 = 0.05$

Figure 13. The images on the far left are simulated TEM images with a single point defect. The point defects are shown in the enlarged window. A sliding window is applied to the TEM images using both the CutPaste and PCA-CNN models. The heatmaps generated by each model are shown to the right of the TEM image. The pixel values in the heatmap represents the average $P(\text{defect})$ at that pixel.

defect to a large TEM image. We then apply a sliding window with a fixed stride and analyze each window using the PCA-CNN and CutPaste. After analyzing each window, we predict the location of the defect to be the pixel location with the highest average CNN output, or $P(\text{defect})$. We compare the predicted defect location with the actual location and determine whether the defect was accurately localized. Figure 13 shows an example of a TEM image with a single antisite defect and the corresponding heatmap generated by the CutPaste model and PCA-CNN model. In this particular example, the antisite defect is more subtle than the rectangular CutPaste defects used for training the CutPaste CNN model, and we find that the CutPaste model is unable to locate the antisite defect. In contrast, the PCA-CNN model is able to clearly locate the antisite defect even though it is trained only using random ciruclar defects.

Table 2 reports the localization accuracy of both models for different defect types and with different levels of imaging noise. The results for the central defocus conditions are also reported separately since the central defocus conditions represent the most likely imaging conditions for experimental TEM image. We exclude circular defects from the results since the PCA-CNN model has the distinct advantage of being trained on circular defects. We only use substitution and antisite defects to compare the performance of the two models since they are excluded from the training data for both models.

In comparing the PCA-CNN and CutPaste models, we find that the PCA-CNN model significantly outperforms the CutPaste model in localizing substitution and antisite defects. Specifically, the PCA-CNN model is able to accurately locate 89% or more of substitution and antisite defects when the variance of the additive Gaussian imaging noise is 0.05 or 0.00. In contrast, the CutPaste model locates approximately $20 - 50\%$ of defects. Even when there is no imaging noise and only considering the

Table 2. Accuracy of the CutPaste and PCA-CNN model in locating point defects in the test set images. Table 2a shows the accuracy results when including all images in the test set. Table 2b shows the accuracy results when only the nominal defocus conditions are included.

(a) Defect localization accuracy including all imaging conditions.

Method	Noise	Substitution $n = 560$	Antisite $n = 560$
CutPaste	σ^2=0.00	0.31	0.53
	σ^2=0.05	0.17	0.39
	σ^2=0.10	0.08	0.21
PCA-CNN	σ^2=0.00	0.71	0.86
	σ^2=0.05	0.64	0.90
	σ^2=0.10	0.14	0.75

(b) Location detection accuracy for central defocus conditions, { -6nm, 0nm, +6nm }.

Method	Noise	Substitution $n = 240$	Antisite $n = 240$
CutPaste	σ^2=0.00	0.38	0.52
	σ^2=0.05	0.22	0.38
	σ^2=0.10	0.07	0.18
PCA-CNN	σ^2=0.00	1.00	0.98
	σ^2=0.05	0.89	0.99
	σ^2=0.10	0.23	0.89

central defocus conditions, the CutPaste model is only able to locate 38% and 52% of substitution and antisite defect, respectively. In contrast, the PCA-CNN model achieves near perfect accuracy when there is no imaging noise.

2.4.3 Leveraging Subject Matter Knowledge

Although the CutPaste method achieves state-of-the-art performance on the MVTec benchmark dataset, the PCA-CNN significantly outperforms the CutPaste method in localizing point defects in simulated TEM images. In this section, we explore how the PCA-CNN model differs from the CutPaste model and, particularly, how the PCA-CNN model leverages knowledge about crystalline materials and the phenomena under consideration. Unlike the CutPaste model, the PCA-CNN model is tailored for defect detection in TEM images. Despite its overall simplicity, we show

that leveraging knowledge of the phenomena in the design of the algorithm allows the PCA-CNN to outperform a state-of-the-art general purpose model.

A key characteristic of TEM images of crystalline materials is the repetitive nature of the lattice structure. Figure 3 shows a set of TEM images of GaAs under different imaging conditions. Although the imaging conditions alter the TEM image, in each example we see that there is a repeating lattice structure. In the PCA-CNN model, we leverage this knowledge by choosing to use PCA to generate lower-dimensional reconstructions of a TEM image segment. The PCA model uses a linear transformation to map an image to a lower-dimensional representation. The lower-dimensional representation is used to generate a reconstruction of the TEM image using an inverse linear transformation. The linear nature of the PCA reconstruction is well suited for reconstructing TEM images because there is a consistent repeating pattern which can easily be represented in lower-dimensional form. To illustrate the importance of the linear nature of the PCA reconstruction in the context of TEM images, we consider an alternative, nonlinear reconstruction method. Specifically, we consider using autoencoders as an alternative to PCA.

Autoencoders are a type of neural network that is used to encode an image to a lower-dimensional representation via multiple convolutional layers. The lower-dimensional representation, often referred as the latent representation, is then passed to a decoder which uses multiple deconvolutional layers to generate a reconstruction image. To generate reconstructions of TEM images, we use the autoencoder architecture shown in Figure 14. For training, the input to the model is a 64 by 90 simulated TEM image segment and the output is the same image. Notice that the encoder portion of the model maps the input to a tensor of shape $(8, 5, 16)$. Thus, the original $64 \times 90 = 5760$ pixels are compressed to a representation with $8 \times 5 \times 16 = 640$ values. We train the autoencoder using a binary cross-entropy loss and an adam op-

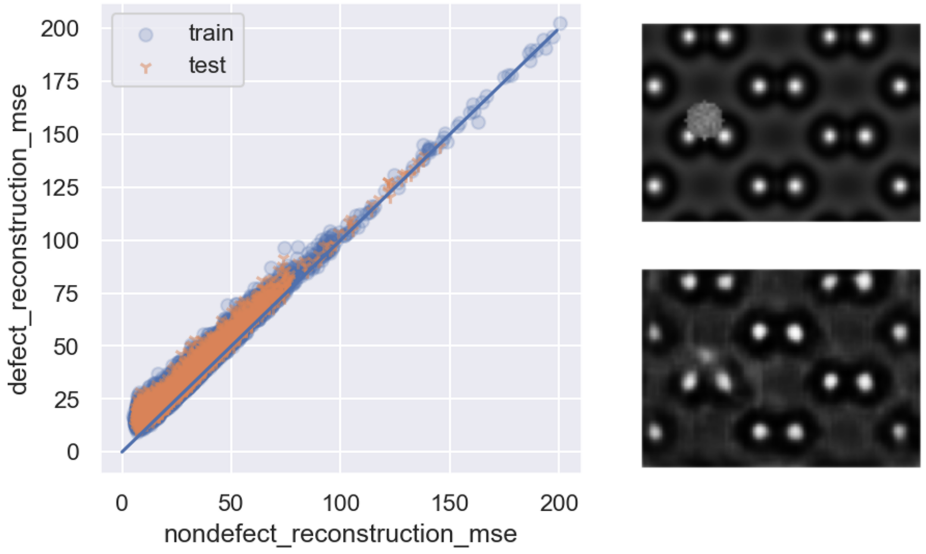
timizer for 2000 epochs. Recall that the purpose of the reconstruction is to subtract the reconstructed image from the original TEM to generate a residual image. The residual image is then passed into a CNN. If the reconstruction can accurately depict the repeating lattice structure in the original TEM image, then the resulting residual image will primarily consist of imaging noise and any point defects.

Layer (type)	Output Shape	Param #
input_2 (InputLayer)	[(None, 64, 90, 1)]	0
conv2d_7 (Conv2D)	(None, 64, 90, 32)	320
max_pooling2d_3 (MaxPooling2	(None, 32, 45, 32)	0
conv2d_8 (Conv2D)	(None, 32, 45, 16)	4624
max_pooling2d_4 (MaxPooling2	(None, 16, 15, 16)	0
conv2d_9 (Conv2D)	(None, 16, 15, 16)	2320
max_pooling2d_5 (MaxPooling2	(None, 8, 5, 16)	0
conv2d_10 (Conv2D)	(None, 8, 5, 16)	2320
up_sampling2d_3 (UpSampling2	(None, 16, 15, 16)	0
conv2d_11 (Conv2D)	(None, 16, 15, 16)	2320
up_sampling2d_4 (UpSampling2	(None, 32, 45, 16)	0
conv2d_12 (Conv2D)	(None, 32, 45, 32)	4640
up_sampling2d_5 (UpSampling2	(None, 64, 90, 32)	0
conv2d_13 (Conv2D)	(None, 64, 90, 1)	289

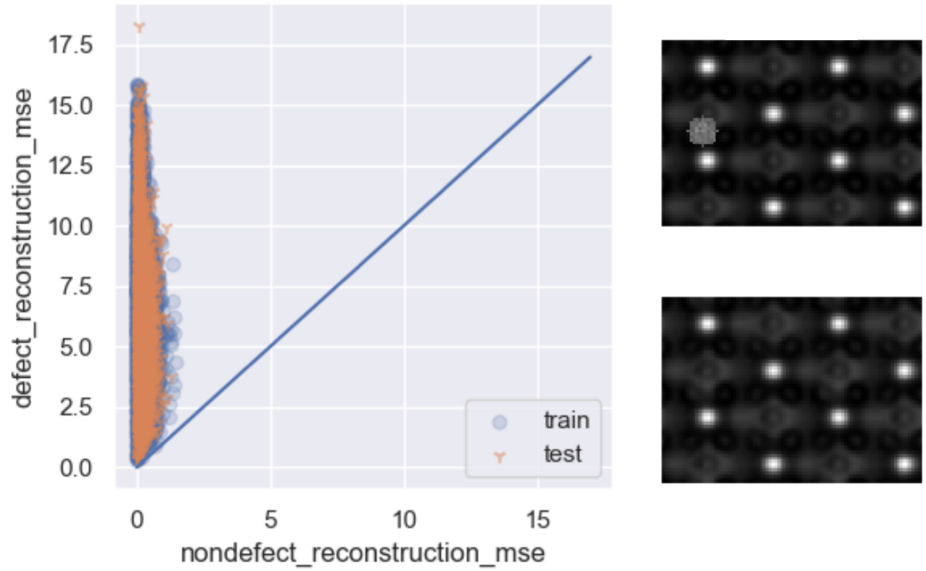
Total params: 16,833
Trainable params: 16,833
Non-trainable params: 0

Figure 14. An autoencoder model is commonly used to generate lower-dimensional reconstructions of the inputs. The first half of the model, known as the "encoder" encodes the input image in a lower-dimensional representation. The second half of the model, known as the "decoder", uses the latent representation to generate a reconstruction that has the same dimensionality as the input image.

However, in addition to accurately reconstructing the repeating lattice structure, we want the reconstruction simultaneously exclude any point defects from the reconstruction. Exclusion of the point defect from the reconstruction ensures that the point defect remains in the residual image. In the PCA-CNN model, we rely on the PCA reconstruction to both reconstruct the repeating lattice and also exclude any point defects from the reconstructed image. Recall, the input to the CNN in the PCA-CNN model is the residual image between the original TEM image and



(a) Using an autoencoder for reconstructions



(b) Using PCA for reconstructions

Figure 15. The scatterplots show the reconstruction error measure by MSE for TEM images with and without defects. Each point in the scatter plot represents a single TEM image in the training set. The x-axis represents the reconstruction MSE when the TEM image has no defects. The y-axis represents the reconstruction SSE when the TEM image has a random circular defects. If all points lie on the line $y = x$, the reconstruction error is the same with or without a defect. If all the points are above the line $y = x$, the reconstruction error is higher when there is a defect. On the right are examples of reconstructions using an autoencoder and PCA.

the reconstruction. If the reconstructed image perfectly reconstructs point defects, then the residual image of a defect image segment and normal image segment will be indistinguishable. Thus, it is important to consider how the PCA model and autoencoder model handle point defects when generating reconstructions. Figure 15 shows a comparison between PCA and autoencoder reconstructions when a point defect is present. We see that point defects can have significant local effects on the autoencoder reconstruction, while the PCA reconstruction does not exhibit any local effects due to point defects. This is explained by the fact that the PCA latent representation is of the form $\mathbf{q_k}^T = \mathbf{q}^T\mathbf{W_k}$ where $\mathbf{q}$ represents a single TEM image stored as a vector and $\mathbf{W_k}$ is the trained PCA transformation matrix. The latent representation, $\mathbf{q_k}$, is then a k dimensional vector where each of the k components is a weighted sum of all the pixel values in the image. Since each component of the latent representation is generated using every pixel value in the image, local point defects are likely to have a minor effect on the latent representation. Autoencoders, in comparison, generate a latent representation using repeated convolutional filters. Each filter considers a local area and generates feature maps by repeatedly apply a small filter to each part of an image. Thus, if our goal is generate reconstructions that do not reconstruct point defects, using a PCA model is likely to outperform an autoencoder. Additionally, PCA models are trained via a closed form solution while training autoencoders are computationally expensive. The PCA used in the PCA-CNN model uses a latent representation with 150 components and takes less than a minute to train with over 5,000 training data images. In contrast, it takes nearly 24 hours to train the autoencoder depicted in Figure 14 for 1000 epochs using the same training data. We can compare the overall performance of the autoencoder and PCA models by computing the reconstruction error for a set of TEM images with and without a defect. Figure 15 shows the reconstruction error with and without a circular defect for each image

in the training and test set. There are two key observations. First, the reconstruction errors, measured in MSE, is nearly an order of magnitude lower using PCA rather than an autoencoder. Notably, the reconstruction error for TEM images using PCA is nearly zero for all imaging conditions. Second, when a defect is added to the image, the PCA reconstruction error substantially higher than the reconstruction error of the same TEM image without a defect. In contrast, when a defect is present, the autoencoder reconstruction error is only slightly higher than the reconstruction error of the same image without a defect. This suggests that the PCA reconstructions are generally unaffected by the presence of a defect, hence the higher error. The autoencoder reconstruction errors are much closer for images with and without a defect. This suggests that the defect influences the reconstruction. Figure 15, upper right, shows an example of an autoencoder reconstruction and reconstruction is clearly affected by the defect. The PCA reconstruction (Figure 15, lower right), in comparison, is seemingly unaffected by the location of the point defect.

2.4.4 Experimental TEM Images

In prior sections, we used simulated TEM images to train and evaluate defect localization methods. In practice, the goal is to use machine learning models to locate point defects in experimental TEM images. In this section, we analyze a single experimental TEM image of a non-GaAs material (Figure 16) that is known to be free of defects. Experimental TEM images have several notable differences from simulated TEM images. First, the imaging noise in a TEM image can be inconsistent and localized to specific regions of an image. Simulated noise is consistent throughout an image. Second, the lattice structure in a simulated image consists of a perfectly repeating pattern such that the spacing between atoms is constant throughout the image. In experimental images, the spacing between atoms can be distorted due

to various environmental factors such as vibrations or slight imperfections in the crystalline material. For these reasons, identifying point defects in experimental TEM images present unique challenges. Furthermore, while the density of point defects in a crystalline material can be controlled during the growing process, the precise location of point defects are unknown. Thus, it is not possible to generate a labeled data set.

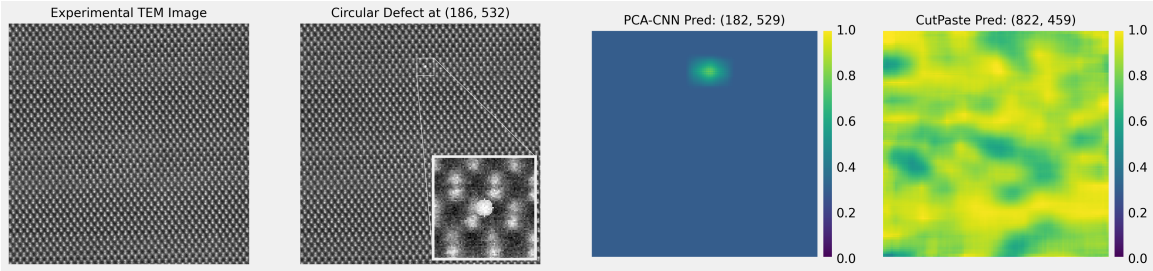


Figure 16. We analyze an experimental TEM image (left) of a non-GaAs material. A random circular defect is added to the image (second from left). The PCA-CNN and CutPaste models are used to localize the added defect. The PCA-CNN model clearly identifies the defect with most of the image having a low P(defect). The CutPaste model fails to recognize that the majority of the image is free of defect. Thus, it struggles to identify the point defect.

To apply the PCA-CNN model on an experimental image, we first retrain the PCA portion of the model using image segment from the experimental TEM image. Intuitively, we use the PCA to learn and reconstruct the repeating "pattern" of the crystalline material so that we can subtract it away. The PCA model can be retrained on an experimental TEM image regardless of whether there are point defects or not because, in practice, the density defect is low and the vast majority of the image is defect free. Retraining the PCA with 200 components can be done in a matter of seconds. For the CNN portion of the PCA-CNN model, we continue to use the CNN that was trained using simulated GaAs images and circular defects in section 2.3. Using the retrained PCA and the pre-trained CNN, we generate a heatmap that computes the P(defect) for each pixel in the experimental image. Figure 16 provides n an example of a PCA-CNN heatmap of the experimental TEM image with a single

circular defect added to the image. Despite the inconsistent noise in the experimental image, the heatmap exhibits low-variance in the P(defect) computed for areas of the image that do not have a defect and clearly identifies the location of the point defect. It appears that the retrained PCA model allows for the PCA-CNN to adapt to the non-GaAs material despite only being trained on simulated images of GaAs.

The CutPaste model is also applied to the experimental image. As shown in Figure 16, the CutPaste heatmap is unable to determine that the majority of the image is defect free. Rather, the CutPaste model predicts that the majority of the image has a P(defect) > 0.50 and is, therefore, unable to clearly locate the circular defect. The CutPaste model is trained on simulated images of GaAs and, unlike the PCA-CNN model, does not incorporate a reconstruction or residual image. Since the experimental image is a non-GaAs material, the atomic columns in the material differ in size and orientation from that of GaAs. These differences may be the reason why the CutPaste model predicts a high probability of defect for the entire image.

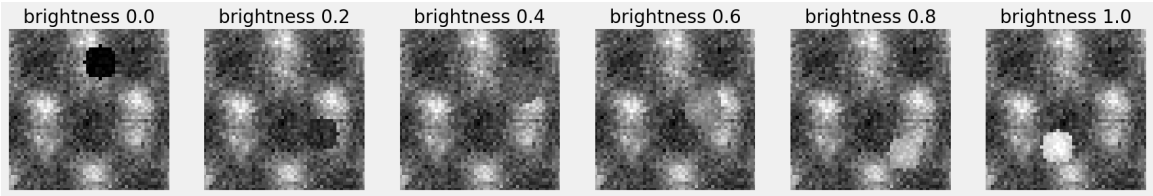


Figure 17. Circular defects of varying brightness are added to the experimental image. Varying the brightness level helps determine the sensitivity of the PCA-CNN and CutPaste methods to point defects in experimental images.

We develop a more comprehensive understanding of the ability of PCA-CNN and CutPaste models by adding randomly located circular defects of varying brightness to the experimental image and assessing whether the two models are able to locate the defect. Specifically, we use six different levels of brightness (Figure 17). For each brightness level, we insert a circular defect in 200 random locations and assess whether the two models are able to locate the defect. Table 3 shows the localization accuracy

for each model at varying levels of defect brightness. The PCA-CNN model is able to locate 100% of defects that are at maximal brightness. The CutPaste model, in contrast, locates less than 20% defects at the maximum brightness level.

Table 3. Accuracy of the CutPaste and PCA-CNN model in locating circular point defects in an experimental image. Circular defect of varying brightness are added to an experimental image.

Defect Brightness	PCA-CNN ($n = 200$)	CutPaste ($n = 200$)
0.0	77.0 %	12.5 %
0.2	59.0 %	13.5 %
0.4	13.0 %	8.5 %
0.6	28.0 %	7.0 %
0.8	80.0 %	17.0 %
1.0	100.0 %	19.5 %

Lastly, we add a subtle and more realistic defect to the experimental image and assess whether PCA-CNN model is able to locate the defect. As in Section 2.4, we consider an antisite defect where two atoms are in swapped locations. Figure 18 displays the defect and the corresponding heatmaps. The PCA-CNN heatmap clearly locates the antisite defect although the $P(\text{defect})$ at the defect is only slighter higher than the surrounding regions (note the scale of the colormap). Given the subtle nature of the antisite defect and the fluctuations in imaging noise throughout the experimental image, it is a notable achievement that the PCA-CNN is able to locate the defect.

2.5 Conclusion

In this chapter, we present a novel method, PCA-CNN, for localizing defects in TEM images of crystalline materials. The PCA-CNN method is a self-supervised method that can be trained entirely on TEM images that are free of defects and exhibits strong performance on simulated data. The ability to train a defect localization

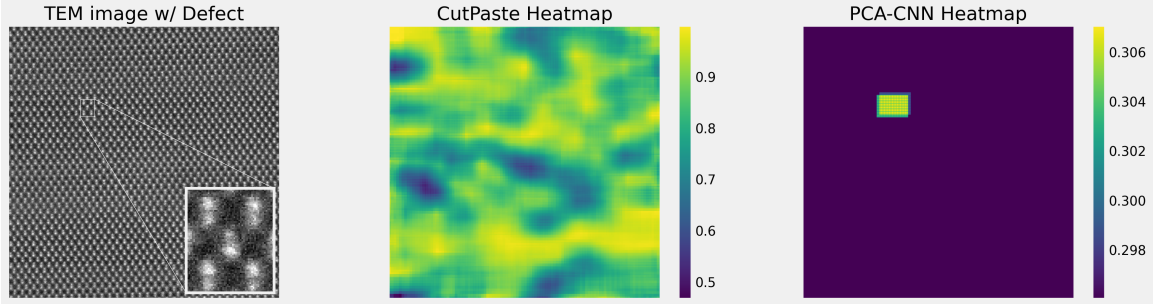


Figure 18. We add a single antisite defect to the experimental image by swapping the locations of two atoms. The defect is shown at the center of the zoomed, inset image. The PCA-CNN heatmap clearly locates the defect.

method without labeled examples of defects represents a novel methodological contribution. Notably, the design of the PCA-CNN method leverages knowledge about point defects in crystalline materials. We show that the tailored design of the PCA-CNN method allows it to outperform CutPaste, a state-of-art, general-purpose defect localization method. Furthermore, we demonstrate the flexibility and generalization performance of the PCA-CNN model by applying it to an experimental image of an unknown crystalline material.

III. Contribution 2: Influence of Pilot Attributes on Fighter Mishap Rates via Bayesian Analysis

3.1 Overview and Motivation

Military aviation safety has received significant attention in recent years due to numerous high-profile mishaps. These mishaps have involved a wide range of aircraft from Army helicopters to next-generation Air Force fighter jets. From 2013-2018, United States (U.S.) Department of Defense (DOD) aircraft accidents from non-combat operations resulted in over $9.81B in damages, 157 aircraft destroyed, and 198 deaths [23]. In response, Congress commissioned the National Commission on Military Aviation Safety (NCMAS) to assess and identify causes contributing to military aviation mishaps. In its report, the commission identifies pressing issues pertaining to pilot and maintainer experience, maintenance logistics, inconsistent funding, and data collection. Similar to many prior studies on military aviation safety, the NCMAS report relies heavily on site visits and interviews rather than quantitative analysis. Notably, the commission highlights that deficiencies in data quality have, thus far, hindered the military community from using data analysis to improve aviation safety.

Existing research on the relationship between pilot behaviors and military aviation mishap rates often relies on qualitative analyses of mishap reports due to the limited availability of quantitative data [24, 25, 26, 27, 28, 29]. The few available quantitative studies focus on the relationship between mishap rates and quantifiable attributes of an aircraft such as aircraft age, number of engines, and mission type [30, 31, 32]. Given that prior qualitative studies, including the NCMAS report, have concluded that flight mishaps can often be attributed to pilot error, a quantitative study of mishap rates that that does not account for pilot attributes or characteristics is incomplete.

Motivated by this shortfall and the substantial consequences associated with mil-

itary aviation mishaps, this study quantifies attributes of pilot communities and then models the relationship between pilot attributes and rates of class A, B, and C flight mishaps [1], hereafter referred to as high class mishaps (HCMs). We focus our analysis on the six distinct fighter aircraft types within the U.S. Air Force (USAF), each referred to by a unique mission design series (MDS), using DOD administrative and accident data from 2008-2020. After quantifying pilot attributes, we model the relationship between pilot attributes and the rate of HCMs using a Bayesian regression framework. We find evidence of a meaningful relationship between pilot attributes and HCM rate. Specifically, we find evidence that a 0.1 standard deviation increase in flying hours over the past 12 months is associated with a 1.4% decrease in HCM rate. This is consistent with the NCMAS's conclusion that a reduction in flying hours is likely to increase the risk of mishaps. Additionally, we find that pilot communities with a 0.1 standard deviation higher proportion of instructor pilots (IPs), distinguished graduates from commissioning source (DGs), and advanced academic degrees are associated with 1.8%, 2.4%, and 1.4% lower HCM rate.

In addition to these model results, our study makes several notable contributions to the existing literature on military aviation safety. First, our analysis of mishap data from 2008-2020 reveals that cost estimates of many class C mishaps are clustered around their upper cost threshold and may have been altered to avoid classification as class B mishaps. As a result, we choose to focus our analysis on the combined rate of all class A, B, and C mishaps. Our inclusion of class B and C mishaps represents a departure from prior studies that have focused solely on class A or fatal mishaps [30, 31]. Second, we present an analysis of the attributes of pilot communities for each MDS. Specifically, we consider attributes pertaining to flying experience and

[1]Section 3 will formally define and discuss these categories of mishaps. Until that point in the article, it suffices to understand that class A mishaps are more severe than class B mishaps and, in turn, class C mishaps, as they relate to the damage incurred to people or property.

proficiency as well as personal demographic data. These attributes are then weighted by the flight hours flown by each pilot in the MDS community which provides a better understanding of the average pilot in the cockpit. Although prior qualitative studies have found that most serious flight mishaps are attributable to pilot error [23], there have been minimal efforts to quantify the characteristics of military pilot communities. Third, our Bayesian modeling approach combined with the use of predictive projection for feature selection represents a methodological contribution to existing aviation mishap literature. Our analysis begins with non-informative priors. Then we demonstrate how the flexibility of Bayesian priors can be particularly useful in applications, such as in military aviation mishap analysis, where existing qualitative research is extensive. Lastly, our analysis provides quantitative insights to complement the causal findings from the NCMAS report. Based on interviews and site visits, the comprehensive NCMAS report finds that several pilot attributes, such as flying experience and fatigue, have a *causal* effect on mishap rates. By modeling the associative relationship between pilot attributes and mishap rates, we investigate whether there is evidence to support the proposed causal relationships in the NCMAS report.

The remainder of this chapter is structured as follows: Section 2 describes related works in the aviation safety literature; Section 3 provides an overview of USAF fighter aviation and safety policy; Section 4 presents the data and modeling framework; Section 5 provide model results; Section 6 discusses these findings and considers the use of informative priors; and Section 7 concludes.

3.2 Related Works

Non-military aviation is categorized as either General Aviation (GA) or commercially related. GA includes all civilian aviation apart from operations involving paid

passenger transport. From 1984-2017, GA flying accounted for 94% of all civil aviation deaths [33]. To contextualize these accident rates, Sobieralski estimates the GA accident rate from 1990-2009 to be over 13 times greater than the motorcycle accident rate, and the fatality rate to be over 10 times greater [34].[2] Motivated by these high accident and fatality rates, Bazargan analyzed the impact of GA pilot characteristics on the likelihood of being in an accident [35]. The authors discovered that a pilot's age and gender do not impact the probability of being in a GA accident. However, Bazargan shows pilots' experience plays a substantial role in predicting accidents in which individuals with less experience are more likely to make errors that lead to an aviation mishap.

Although the GA accident research provides important insight into aviation mishaps, the structure of commercial aviation and its respective accident rates are more comparable with military aviation.[3] Barnett [38] finds that, in commercial aviation, the death risk per flight fell by more than half from 2008–2017 compared with the previous decade. Much analysis has also been conducted on human-factors such as fatigue in order to explain changes in aviation mishaps [39, 40]. Additionally, there has been some important work by Haunschild [41] on how the commercial airlines have learned from previous accidents, leading to fewer aviation mishaps over time. This research complements other work in the transportation sciences that show how organizations can learn from previous mistakes [42, 43, 44]. Nonetheless, there is a lacuna in this literature of analyzing the relationship between pilot attributes and accident rates. In contrast, there is a wide literature regarding the relationship between driver demographics and vehicle accident rates. Driver demographics such as gender,

[2]Sobieralski also estimates the costs of these deaths range from $1.6 to $4.6 billion using models on the statistical value of a human life and the willingness-to-pay approach.

[3]Additionally, research by Oster and Rios [36, 37] show increases in aviation safety provide significant positive effects for the global economy, yet few studies have examined the characteristics of the pilot communities that contributed to the positive trends.

age, and driving experience has been incorporated in numerous traffic safety studies [45, 46, 47, 48, 49, 50, 51, 52].

Military aviation safety research can be broadly categorized as either qualitative or quantitative. Qualitative studies are generally based on an analysis of mishap investigation reports or extensive surveys with aircrew and maintenance personnel. The recent report from the NCMAS is a comprehensive, largely qualitative, analysis of military aviation safety based on over 200 site visits and numerous interviews. The study proposes a wide range of causal reasons for rising safety mishaps including reduced pilot and maintainer experience, increased operational and personnel tempo, uncertain budget requirements, and data deficiencies [23]. Earlier studies analyze mishap reports for human factors trends and conclude that crew resource management (CRM) is a significant factor in aviation mishaps [26]. Miranda et. al. [24] analyze the role of human factors in a limited number of Naval mishaps and conclude that teamwork failures play a significant role in mishaps. Other examples of qualitative works make similar conclusions about the role of human factors [27, 28, 29]. Whereas these studies are useful for understanding the nuances of aviation mishaps and for forming hypotheses about causal effects, the reliance on manually reviewing individual mishap reports limits the ability to study larger scale trends across various aircraft types.

Related works that focus on the quantitative analysis of military aviation mishaps are less common, but progress has been made in recent years. Pamplona et. al. [31] consider mishap and fatality rates for each MDS separately and fit a distribution to model fatality risk by MDS. Light et. al. [30] model the number of yearly mishaps as a function MDS, aircraft age, and other aircraft characteristics such as whether an aircraft is multi-engine or single-engine. Additionally, the authors find that, similar to commercial mishap rates, military aviation accident rates have been on a downtrend

since 1950, specifically in the number of class A mishaps and destroyed aircraft. However, the NCMAS [23] report highlights that improvements in military aviation safety have not kept up with improvements in commercial aviation since the 1990s. Both of these works primarily capture the difference in mishap rates between different MDS without accounting for pilot attributes. Gaines et. al. [25] focus specifically on the frequency of fatigue as a contributing factor in USAF mishaps. In addition to these recent works, an earlier study models class A mishaps rates as function of cumulative flight hours and concludes that mishap rates generally decrease as each MDS accumulate more flight hours [32]. Similar to the commercial aviation safety literature, there is an opportunity to expand on this research by incorporating data regarding pilot characteristics.

3.3 Background

Pilots in the Air Force are commissioned officers and college graduates who have, in the vast majority of cases, been commissioned via the Reserve Officer Training Corps (ROTC), United States Air Force Academy (USAFA), or Officer Training School (OTS). Upon commissioning as an officer, future pilots are assigned to one of four Undergraduate Pilot Training (UPT) bases: Laughlin Air Force Base (AFB), Texas; Vance AFB, Oklahoma; Columbus AFB, Mississippi; or Sheppard AFB, Texas. At their assigned UPT base, all trainees follow a similar syllabus and fly the T-6A Texan II aircraft during the Primary Phase. The Primary Phase lasts approximately 28 calendar weeks and involves around 87 hours of total flight time in the T-6A [53]. Based on initial performance and preference, UPT students are then separated into one of three different "tracks" during the Advanced phase. Those selected for the fighter and bomber track (i.e., and not the tanker, heavy, or helicopter track) train in an advanced trainer aircraft during the second half of UPT, where they accumulate

around 96 hours in the T-38C [54]. Based on UPT performance, preference, and availability of fighter aircraft, students are selected to become fighter pilots and are matched with one of six MDS: A-10, F-15, F-15E, F-16, F-22, or F-35. MDS types can differ in terms of aircraft age, number of engines, and mission type. Additionally, each MDS is generally assigned to a different set of base locations which influences the geography and climate in which aircraft are flown.

Once assigned to an MDS, fighter pilots will generally fly the same MDS for the majority of their career.[4] Although they fly the same MDS throughout their career, fighter pilots are assigned to a different unit and base every few years [55]. Figure 19 summarizes the early career path for fighter pilots.

The Department of Defense (DOD) Instruction 6055.07 defines various classes of aviation mishaps and the investigation requirements for each class of mishap. In general, the most serious mishaps are class A mishaps followed by class B, C, D, and E. Figure 20 shows the flowchart for determining the classification of a mishap. The Air Force Safety Center (AFSEC) is responsible maintaining safety programs and policies according to DODI 6055.07. In our study, we focus on class A, B, and C mishaps that occur during flight. We provide justification for why we choose to focus on these mishap classes in the following section. Importantly, the NCMAS report and prior studies find that the vast majority of class A, B, and C mishaps that occur during flight are attributable to pilot error [23].

[4]Pilots have the opportunity to return to UPT as instructors later in their career. Additionally, there are career broadening opportunities for fighter pilots to potentially fly other aircraft in different military branches and/or partner nations' air forces. Finally, they may also transition to fly newer aircraft, such as F-35s which were initially unavailable to recent graduates of UPT. Until 2018, the F-35 was only accessible to experienced pilots who would cross-train from a different platform.

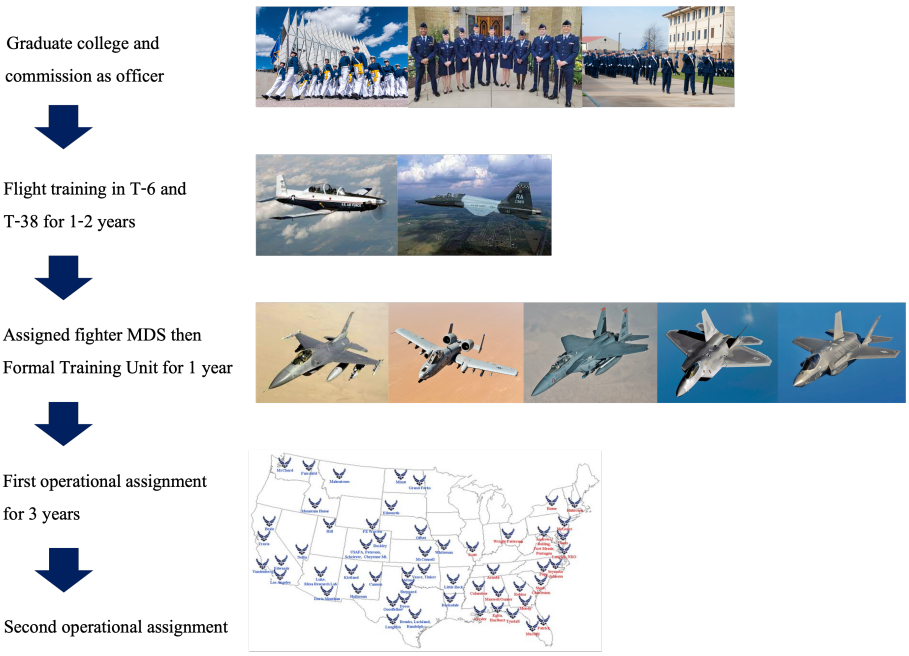


Figure 19. Summary of early career path for fighter pilots in the US Air Force.

3.4 Materials and Methods

3.4.1 Data

For this study, we utilize both aviation mishap records from the Air Force Safety Center (AFSEC) and military personnel records from the DoD. Before presenting the data, we echo the NCMAS report's general concerns about data deficiencies and the effect it has on data analysis. Preparing the safety and personnel data for analysis required significant cleaning due to inconsistencies and missing data. For example, we found that many mishaps were missing important data such as the identification of squadron or unit involved. The lack of unit information precluded us from modeling the relationship between pilot attributes and mishap rates at a unit-level. In the personnel data, we found inconsistent record keeping of flight hours that resulted in pilots "losing" past flight hours that would reappear in later years. We also encountered numerous minor record-keeping issues such as multiple names for the same

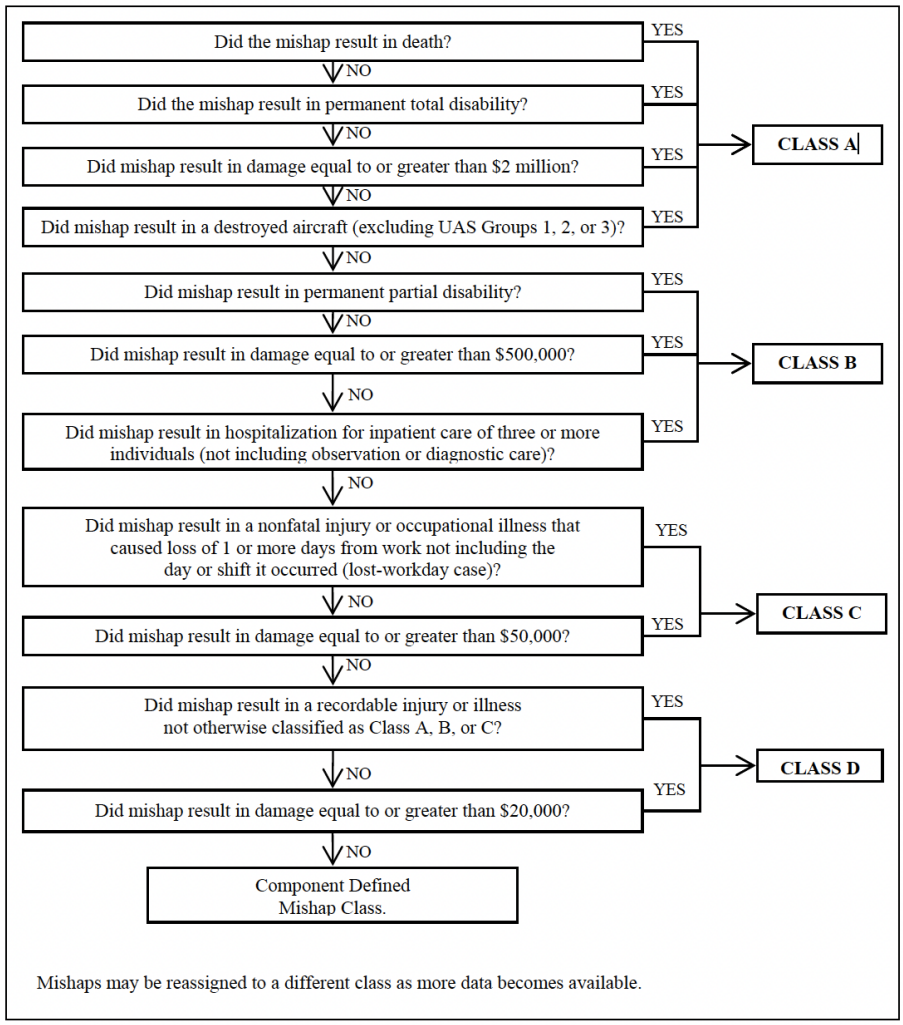


Figure 20. Guide for determining mishap class according to *DODI 6055.07* until 2019 [56]. The cost thresholds were increased in 2020 to $60k, $600k, and $2.5m for class C, B, and A mishaps, respectively.

MDS and for the same unit. Lastly, merging the two datasets presented challenges because the safety and personnel datasets use different nomenclature for MDS types and units (i.e., squadrons, groups, and wings). Improving the quality of data sources and ensuring data consistency across organizations would enable greater use of data analysis to improve aviation safety.

We first present the safety data from AFSEC. AFSEC maintains a record of all mishaps and hazards using the Air Force Safety Automated System (AFSAS). When a safety incident occurs, AFSAS is used to record details on the severity, location, and aircraft involved. Our AFSAS dataset includes mishap data for all Air Force fighter safety incidents from 2007-2020. The dataset includes 31,000 records and, for each record, contains 39 data fields. Figure 21 shows the rate of class A, B, and C flight mishaps from 2007-2020. Although there appears to be a rise in the rate of class A mishaps in recent years, the overall rate of class A, B, and C mishaps remaining relatively steady.

In this study we focus on the total rate of class A, B, and C flight mishaps, or HCMs, for fighter aircraft[5]. While other studies have focused on modeling class A mishaps or fatalities [30, 32], we include class B and C mishaps in our primary response variable for two reasons. First, the NCMAS commission notes that safety leaders across the military consistently suggested that "class C mishaps are potentially the best indicators of elevated risk for more serious mishaps" [23] because there is often a marginal distinction between class A mishaps and less severe mishaps. Thus, trends in class B and C mishaps can serve as an indicator of overall safety risk.

Second, we include class B and C class mishaps in our study because our analysis reveals that a significant number of fighter mishaps had cost estimates just below the threshold between class B and C mishaps. Figure 22 shows the distribution of mishap

[5]We exclude all ground mishaps since they are often related to maintainer error [23]

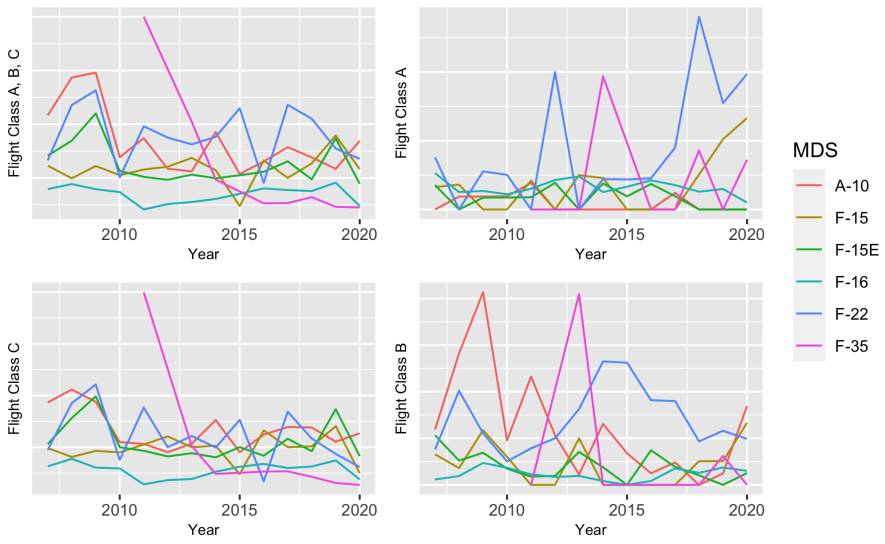


Figure 21. Annual rates, per 100,000 flight hours, for various types of safety events separated by MDS.

costs around the thresholds between class A, B, C, and D mishaps. The distributions of mishap costs around the class C and D threshold seem unaffected by the \$50k threshold. In contrast, the distribution of costs around the class C and B threshold reveals a significant number of mishaps with a cost estimate slightly below \$500k and suspiciously few cost estimates just above the threshold. For instance, from 2007 to 2019, there were 105 mishaps with a cost estimate between \$400k and \$500k in comparison to 33 mishaps with a cost estimate between \$500k and \$600k. If we narrow the range of interest to ±\$50k, there are 60 mishaps with a cost estimate between \$450k and \$500k in comparison to 16 mishaps with a cost estimate between \$500k and \$550k. This anomalous aspect of the distribution suggests that cost estimates may have been influenced by the threshold. We can formally test whether the \$500k threshold affects cost estimates near the threshold using the McCrary sorting test [57]. The McCrary test lends a p-value of 7.3×10^{-5}, and we reject the null hypothesis that there is no sorting or manipulation. Safety personnel could be motivated to manipulate cost estimates because class C mishaps generally have fewer reporting

requirements than class A and B mishaps [23]. The effect of mishap class thresholds on cost estimation behavior warrants further analysis, but in this study, we simply include all class B and C mishaps in our analysis.

Our second primary data source are administrative personnel records provided by the Headquarters Air Force Mission Directorate for Manpower, Personnel, and Services (HAF/A1). This personnel dataset includes quarterly data of all USAF pilots on active duty from 2007-2020. For each active duty member, the data includes detailed career information such as assigned unit, rank, Air Force Specialty Code (AFSC), duty title, time in grade, commissioning source, and flying experience. The flying experience is of particular importance because it provides the flight hours flown on each MDS by each pilot in each quarter. The dataset also includes demographic data such as marital status, household size, level of education, academic major, gender, and race. In total, the dataset contains over 700 data fields per observation.

Using the flying history data, we group pilots by MDS and year to compute the total annual flying hours for each fighter MDS and to quantify personnel characteristics for each MDS pilot community. We quantify the characteristics of an entire MDS pilot community by taking the weighted average of pilot characteristics, where weights are proportional to the number of flight hours on an MDS in a given year. Since we are weighting by the flight hours accrued on a given MDS, the personnel data generally captures the attributes of the pilots who *flew* a particular MDS in a given year rather than average attributes of all pilots in an MDS community in given year. This difference is subtle but important. We focus on the former population sample because pilots who do not accrue flying hours do not affect the observed mishap rate. Taking the weighted average of pilot attributes in each MDS and year community, there are 10 personnel predictors we consider in our analysis. Table 5 describes these 10 predictors and provides the standard deviation of each. The predictors are stan-

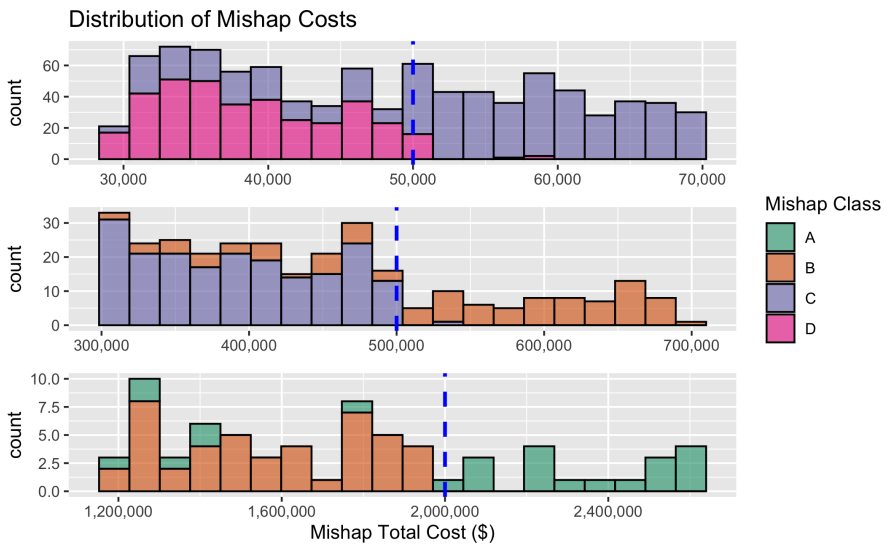


Figure 22. Distribution of mishap costs near the cost thresholds for class A, B, and C mishaps from 2008-2019. The cost thresholds are denoted by vertical dashed lines. The class designation of each mishap is indicated by color.

dardized prior to model fitting so the standard deviations are necessary for model interpretation. The mean values of the predictors are redacted because the operation nature of the data precludes its publication.

We focus on these 10 personnel characteristics because they align with the conclusions of the NCMAS report and other prior, qualitative studies. Based on site visits and interviews with pilots the NCMAS finds that a reduction of average pilot flight hours and proficiency in recent years has caused an elevated levels of risk. We include x_{exp}, x_{lag}, x_{age}, and x_{ip} to capture pilot characteristics related to flying experience and proficiency. Additionally, we include x_{dg} and x_{edu} because the rate of distinguished graduates [6] or graduates degrees may also be related to pilot proficiency. The commission also found that increased distractions such as additional duties and fatigue can increase risk. Accordingly, we include x_{ta}, x_{spouse}, and x_{child} to capture

[6] In general, an officer receives DG recognition at the time of commissioning if they graduate in the top 10% of their program. For USAFA and ROTC graduates, DG status depends on academic, physical fitness, and military leadership performance during a cadet's undergraduate career.

Table 4. For each year from 2008-2020, we consider 10 predictors to characterize each MDS pilot community. The predictors are standardized prior to model fitting. Standard deviations for each predictor are provided to assist in interpreting model results.

Predictor	Description	Std Dev
x_{exp}	career flight hours	181.41
x_{lag}	flight hours in past 12 months	26.89
x_{age}	age	1.25
x_{child}	percent with child	0.09
x_{spouse}	percent married	0.05
x_{ip}	percent who are currently IP	0.11
x_{dg}	percent who are DG from commissioning source	0.06
x_{edu}	percent who have completed advanced academic degree	0.15
x_{ta}	percent who are currently using TA	0.09
x_{afa}	percent graduated of USAFA	0.05

non-flying related commitments and fatigue. Pilots who are using tuition assistance (TA) are assumed to be pursuing a master's degree while still flying full-time. While we chose these personnel factors because we believe each of them could potentially have a causal effect on mishap rates, we emphasize that our inference methods are focused solely on quantifying and identifying associative relationships. We address this matter in further detail in our discussion of the results.

Figure 23 illustrates the trend of six personnel-related characteristics over time. There are several trends worth highlighting. First, the average career flight hours has been steadily decreasing across all MDSs. Note that these are weighted averages, not the simple average across all pilots in an MDS community. Thus, this downward trend indicates that a consistently increasing number of flights hours are being accrued by pilots with fewer career flight hours. Interestingly, the weighted average flight hours in the prior 12 months has remained relatively steady. This difference could be result from a combination of experienced pilots leaving the Air Force and more flight hours being allocated to newer pilots. Second, we see clear evidence of Air Force policy changes such as reduced flying hours in 2013 due to sequestration ([58])

and a drastic drop in TA usage after 2014 when the Air Force removed graduate degrees from promotion reviews[7]. Lastly, we notice that MDS pilot communities can differ substantially. For example, the average career flight hours of F-35 pilots was significantly higher than other MDS pilot communities from 2010 to 2017, and the F-15E community has a higher percentage of pilots who are DGs than the F-16 community. The F-35 is noteworthy because it is the newest Air Force fighter. The training program for F-35 pilots was opened in 2013 and, for most of the 2010's, there were relatively few F-35 pilots in the Air Force [59]. Given the relatively small size and unique characteristics of the F-35 pilot community, we exclude F-35 data for the majority of our subsequent analysis.

Figure 24 shows the correlations between personnel predictors in our data. As expected, certain pairs of predictors are highly correlated, either positively or negatively. For example, age and percent of pilots who are IPs are highly correlated since most pilots become IPs at similar points in their career. We address multicollinearity during our feature selection process by using predictive projection [60].

In addition to the personnel predictors, we also account for MDS type in our models. MDS type is an important predictor because it helps address many issues with unobserved heterogeneity. Each MDS differs with respect to aircraft age, mission type, and number of engines. Additionally, different MDS types are assigned to bases at different geographic locations. Thus, including an MDS predictor controls for a wide range of non pilot-related factors that could influence mishap rate and allows for more accurate estimation of the association between mishap rate and attributes of each MDS pilot community. We use five dummy variables in our models to account for the six MDS types. While dummy variables is a simple method for addressing unobserved heterogeneity, it has been shown to lead to consistent estimates of the

[7]According to Chesney [2022], the variation in education investment behavior of Air Force Officers was mostly the result of promotion policy reforms.

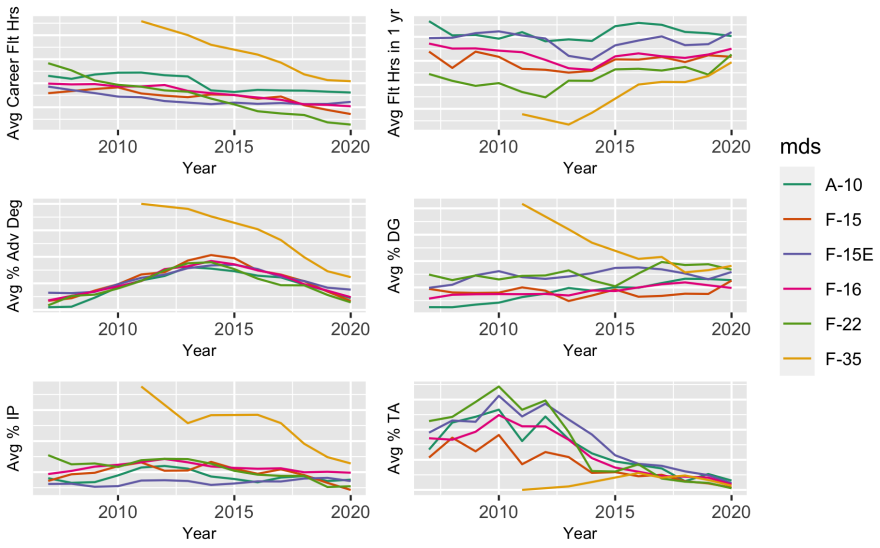


Figure 23. Air Force Personnel data is used to group pilots by MDS within each year. These plots show annual characteristics of each MDS pilot community over time. The upper-left plot shows the total hours flown on each MDS. The remaining plots show average career flight hours, service time, percent distinguished graduate (DG), percent with children, and percent with an advanced academic degree within each MDS pilot community.

coefficients of the remaining predictors [61]. For notational convenience, we refer to the set of MDS dummy variables as x_{mds} in Table 5.

The training dataset, excluding the F-35, consists of 70 observations (6 MDSs over 14 years) wherein each observation represents one year of an MDS pilot community. For each observation, the response variable is the yearly HCM rate and the predictors are pilot attributes and MDS type.

3.4.2 Modeling Framework

Our goal is to model the relationship between HCM rate and personnel factors associated with each MDS pilot community. We model mishap rate using both a Bayesian Poisson and negative binomial (NB) generalized linear model (GLM) with a

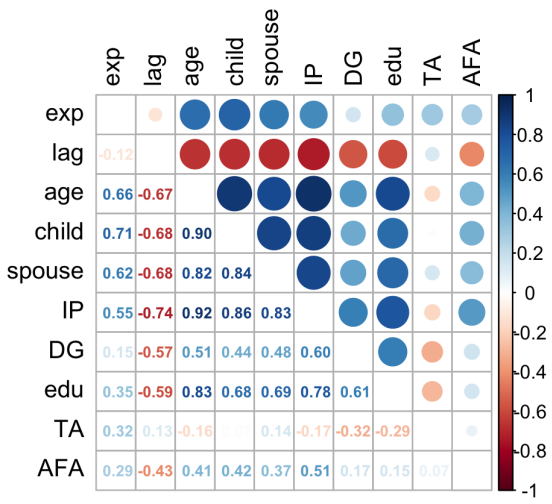


Figure 24. Correlations between pilot community attributes.

log link[8]. To model the mishap rate, we use the mishap count as the response variable and include the log of the MDS flight hours as an offset variable. Specifically, we define the mishap rate, $\frac{y_i}{h_i}$, where y_i is the event count and h_i is the number of known hours flown for observation i. Using a Poisson generalized linear model (GLM) with a log-link, the mishap rate is then modeled as

$$y_i \mid \lambda_i \overset{ind}{\sim} \text{Poi}\left(\lambda_i\right)$$
$$\lambda_i = e^{\ln h_i + \beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j}}$$

where $\ln\left(h_i\right)$ is an offset with fixed coefficient of 1. The NB is similar except for the inclusion of an additional dispersion parameter. For all models, we assume weakly informative Gaussian priors on the coefficients[9] and, for the negative-binomial model, an exponential prior on the reciprocal dispersion term.

[8]Poisson and negative binomial GLMs are both used to model count data. However, the Poisson GLM is commonly critiqued due to the equal conditional mean and variance assumption. As a result, the negative-binomial GLM is often recommended for overdispersed data. Although this line of reasoning is prevalent, it has been shown that the Poisson GLM is robust to violations of the equal mean and variance assumption. Moreover, the negative-binomial GLM has undesirable robustness properties that depend on the nature of the overdispersion [62, 63, 64]. We fit both models and compare performance using ELPD_{loo}.

[9]We use default priors from the R package `rstanarm`

We use expected leave-one-out cross-validation (loo) log pointwise predictive density (ELPD) as the primary metric for model comparison. The ELPD_{loo} metric is defined as

$$\text{ELPD}_{\text{loo}} = \sum_{i=1}^{n} \log p\left(y_i \mid y_{-i}\right), \tag{1}$$

where

$$p\left(y_i \mid y_{-i}\right) = \int p\left(y_i \mid \theta\right) p\left(\theta \mid y_{-i}\right) d\theta \tag{2}$$

is the leave-one-out predictive posterior density given the data without the ith data point([65]. A higher ELPD_{loo} implies that the fitted model is, on average, placing higher density on the left-out, observed response. Additionally, the ELPD_{loo} metric depends on the quality of the predictive uncertainty calibration as well as that of the point predictions.

Although ELPD_{loo} is utilized to compare models such as the Poisson and NB GLMs, adopting the same metric for feature selection can lead to overfitting [60]. Therefore, we use Bayesian predictive projection for feature selection. The use of predictive projection for feature selection in Bayesian regression has a long history [66], but it has received renewed attention in recent years. Recent works [60, 67] show that predictive projection feature selection outperforms alternative methods such as sparsifying priors for identifying relevant predictors, estimating uncertainty, and improving predictive performance. To summarize, the predictive projection first requires a "reference" model, which often includes all available predictors. After fitting a reference model, the goal is to project the reference model to numerous candidate submodel models (i.e., models with fewer predictors) and find a submodel that achieves comparable performance to the reference model. For each candidate submodel, the

parameters of the submodel are determined by minimizing the Kullback-Leibler (KL) divergence between the predictive posterior of the reference model and that of the submodel. Specifically, the parameters for a submodel, π, are determined via

$$\boldsymbol{\theta}_{\pi} = \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathrm{KL}\left(p\left(\tilde{y} \mid \boldsymbol{\theta}_{*}\right) \| p(\tilde{y} \mid \boldsymbol{\theta})\right)$$

where ($p(\tilde{y} \mid \boldsymbol{\theta}_{*})$ and $p(\tilde{y} \mid \boldsymbol{\theta})$ are the predictive posteriors of the reference model and submodel, respectively. Each submodel is fit and assessed multiple times using cross-validation. This cross-validation scheme has been shown to mitigate selection bias issues. Lastly, we select a submodel that achieves comparable performance to the reference model. We use the R package `projpred` to implement predictive projection. The predictive projection method is explained in detail in Piironen2020ProjectiveSelection

The ability to use predictive projection for feature selection is a key benefit of leveraging a Bayesian modeling framework. Additionally, Bayesian models allow for more intuitive interpretation of uncertainty estimates on the parameters as well as straightforward uncertainty propagation. While we choose to use noninformative priors in our analysis, the flexibility of Bayesian priors provides an elegant way to incorporate the findings from existing qualitative studies. In general, Bayesian generalized linear models with noninformative priors result in effect size estimates that are similar to frequentist generalized linear models [68] [10]

In total, we fit and compare five model specifications, M1-M5. For each of the five specifications, we fit both a Poisson and NB regression model with an offset. For models M1-M4, we exclude F-35 data from the training data. We incorporate F-35 data for M5. Model M1 has no predictors and serves as a baseline. Model M2 uses only MDS as a predictor. Model M2 is motivated by prior research which

[10] For comparison, a frequentist Poisson GLM model with robust standard errors was also fit. The results are consistent with the Bayesian GLM results.

has shown that mishap rates can vary greatly for different MDS types [30, 31, 32]. Different MDS types can have differing number of engines and various other physical differences so it is reasonable to expect mishap rates to differ by MDS. Given this context, our research focuses on whether the addition of pilot attributes improves the predictive performance of the model. Thus, Model M3 includes the all available personnel factors in our data set as listed in Table 5. Model M4 includes subset of the available personnel factors. The predictors included in M4 are based on the results of applying predictive projection with M3 as a reference model. We then fit M5 with F-35 data included and the same set of features used in M4.

3.5 Results

Model M1 only includes an intercept and serves as a baseline. As we add additional predictors, we can determine whether the newly added predictors meaningfully improve upon this baseline model. Model M1, results in a ELPD_{loo} of -360 and -257.2 using Poisson and NB GLM, respectively. In general, a higher ELPD_{loo} suggests better predictive performance on out-of-sample data, so we conclude that the negative-binomial is a better fit. The next model specification, M2, includes MDS as the only predictor. Both the Poisson and negative-binomial GLM model improve substantially with ELPD_{loo} increasing to -244 and -234, respectively. The increase

Table 5. Five models are fit using different sets of personnel predictors and data. A Poisson and NB GLM is fit for each of the five models.

Model	Response	Personnel Predictors	Data
M1	y_{HCM}	–	w/o F-35
M2	y_{HCM}	x_{mds}	w/o F-35
M3	y_{HCM}	$x_{\text{mds}}, x_{\text{exp}}, x_{\text{lag}}, x_{\text{age}}, x_{\text{ip}}, x_{\text{child}}, x_{\text{spouse}}, x_{\text{ta}}, x_{\text{dg}}, x_{\text{edu}}, x_{\text{afa}}$	w/o F-35
M4	y_{HCM}	$x_{\text{mds}}, x_{\text{lag}}, x_{\text{ip}}, x_{\text{dg}}, x_{\text{edu}}, x_{\text{ta}}$	w/o F-35
M5	y_{HCM}	$x_{\text{mds}}, x_{\text{lag}}, x_{\text{ip}}, x_{\text{dg}}, x_{\text{edu}}, x_{\text{ta}}$	w F-35

in ELPD_{loo} suggests that MDS type has a meaningful relationship with the rate of HCMs.

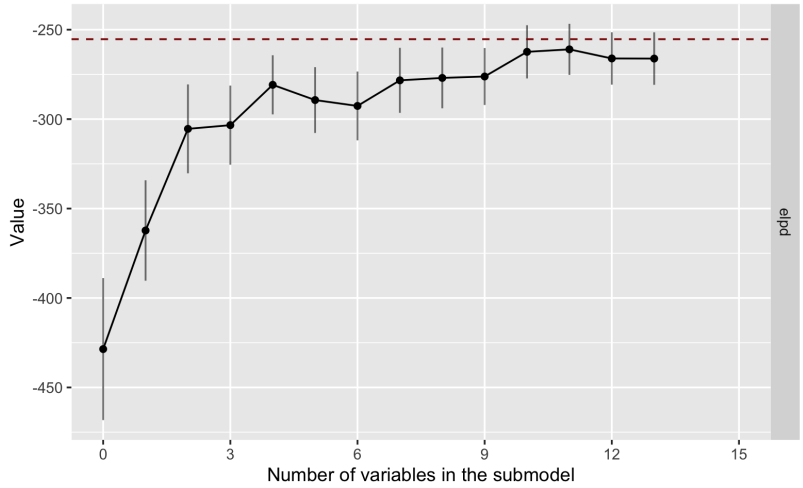


Figure 25. The performance of submodels with increasing numbers of variables. The dotted line represents the performance of the reference model.

In model M3, we include all personnel factors as predictors in the regression. The fitted Poisson and NB models result in ELPD_{loo} of -228.2 and -234.1, respectively. The personnel features do not appear to improve the cross-validated performance of NB regression, but they do lead to a slight improvement in the Poisson regression. The inclusion of all personnel predictors may result in overfitting, so we perform feature selection using predictive projection with model M3 Poisson GLM as the reference model. This process results in Figure 25, which shows the performance of candidates submodels as the number of predictors increases. Table 6 reports the set of predictors associated with each submodel. We default, we decide to include all MDS types as predictors. Next, we select the select a submodel such that the submodel's performance is within one standard error of the reference model. Thus, we choose to include personnel predictors up to IP percentage because the submodel of size 7 is within nearly one standard error of the reference model Piironen2020ProjectiveSelection. This process results in only MDS, TA percentage, IP percentage, DG percentage, and advanced

Table 6. TEST TEST

Num Variables	Variable	ELPD	ELPD Std Error
0	intercept	-428.56	39.91
1	lag	-362.25	28.22
2	F-16	-305.44	24.98
3	edu	-303.36	22.24
4	TA	-280.82	16.61
5	F-15	-289.34	18.49
6	DG	-292.63	19.33
7	IP	-278.32	18.26
8	exp	-276.96	17.07
9	F-15E	-276.18	16.00
10	AFA	-262.38	14.96
11	child	-261.00	14.36
12	age	-266.08	14.71
13	F-22	-266.15	14.75

academic degree percentage being included in model M4. Using this subset of personnel predictors yields Poisson and NB regressions with ELPD_{loo} of -224.9 and -230.4, respectively. A summary of the ELPD_{loo} for each of the fitted models are shown in Table 7.

After fitting models M1-M4, we further investigate whether the inclusion of personnel predictors improves predictive performance in comparison to only using MDS as a predictor. The M2 NB model achieves an ELPD_{loo} of -234 and outperforms the M2 Poisson model. We compare the M2 NB model with the M4 models. Using the R package `loo` implementation for model comparison, we find that the M4 Poisson yields an ELPD_{loo} that is 9.1 units higher than the ELPD_{loo} of the M2 NB model. This difference has a standard error of 6.7. The M4 NB has an ELPD_{loo} that is 3.8 units higher than the ELPD_{loo} of the M2 NB model. This difference has a standard error of 2.5. Note that these standard errors are for the estimated difference in ELPD_{loo} between two models and are not equivalent to the standard errors in Table 7 [69]. Given the differences in ELPD_{loo} values are approximately 1.4 and 1.5 stan-

	Poisson		Negative-Binomial	
	ELPD_{loo} (SE)	p-loo (SE)	ELPD_{loo} (SE)	p-loo (SE)
M1	−360.0(38.9)	7.7(1.7)	−257.2(7.4)	1.5(0.3)
M2	−244.0(14.7)	10.4(2.3)	−234(5.7)	3.3(0.6)
M3	−228.2(8.3)	18.4(2.8)	−234.1(4.3)	7.1(0.8)
M4	−224.9(9.2)	12.9(2.3)	−230.3(4.6)	4.9(0.7)
M5	−246.3(10.4)	14.6(2.6)	−250.1(5.9)	6.3(0.9)

Table 7. We use ELPD_{loo} as the primary metric for comparing models. Higher values of ELPD_{loo} indicate better performance. Another metric, p-loo, is also reported. P-loo is the difference between ELPD_{loo} and the non-cross-validated log posterior predictive density. Higher values of p-loo indicate indicate that predictive performance on future observations is more difficult.

dard errors from zero, respectively, there is evidence that the addition of personnel factors improves the out-of-sample prediction of HCM rates for both the Poisson and NB GLMs. We can visualize the cross-validated performance of each of the models using several posterior checks. Figure 26 shows the distribution of LOO and standard residuals for models M1-M4. We observe that the spread of the LOO residuals progressively narrows. This trend supports the conclusion that adding pilot personnel predictors improves the predictive performance of a model relative to an alternative including only includes MDS as a predictor.

Figure 27 depicts the distribution of the predictive posterior mean and standard deviation in comparison to the mean and standard deviation of the observed HCM rate. When comparing the M4 Poisson and NB models, the predictive posterior of the M4 Poisson is better aligned with the observed sample statistics. In particular, the M4 NB model generates predictive posterior samples that are overdispersed relative to the observed HCM rates. Thus, we focus on the posterior results of the M4 Poisson model. The posterior estimates for all models are summarized in Table 8, and the posterior distributions for the M4 Poisson parameters are shown in Figure 28. We notice that F-16s and F-15s are associated with a decreased HCM rate relative to the

Table 8. Summary of modeling results. Point estimates are the median of the posterior distributions. The standard error estimates are shown in parentheses.

	Dependent variable				
	HCM rate				
	M1 NB	M2 NB	M3 Poi	M4 Poi	M5 Poi
(Intercept)	-7.7	-7.30	-7.55	-7.47	-7.38
	(0.0)	(0.10)	(0.17)	(0.11)	(0.11)
F-15		-0.36**	-0.39*	-0.52***	-0.57***
		(0.15)	(0.21)	(0.14)	(0.14)
F-15E		-0.29**	0.17	-0.06	-0.08
		(0.14)	(0.21)	(0.11)	(0.12)
F-16		-0.88***	-0.66***	-0.88***	-0.95***
		(0.14)	(0.15)	(0.19)	(0.11)
F-22		0.06	0.39	0.17	0.03
		(0.15)	(0.30)	(0.21)	(0.21)
F-35					-0.72***
					(0.29)
exp			0.23		
			(0.19)		
lag			-0.15	-0.15*	-0.16*
			(0.09)	(0.09)	(0.09)
age			0.32*		
			(0.17)		
child			-0.04		
			(0.12)		
spouse			-0.13		
			(0.09)		
IP			-0.38***	-0.19*	-0.09
			(0.18)	(0.10)	(0.09)
DG			-0.13	-0.27***	-0.22***
			(0.10)	(0.07)	(0.07)
AFA			0.03		
			(0.07)		
edu			-0.16*	-0.15***	-0.17***
			(0.1)	(0.06)	(0.05)
TA			-0.12**	-0.05	-0.05
			(0.05)	(0.04)	(0.04)
Observations	70	70	70	70	78

*90%, **95%, ***99% credible intervals exclude 0.0

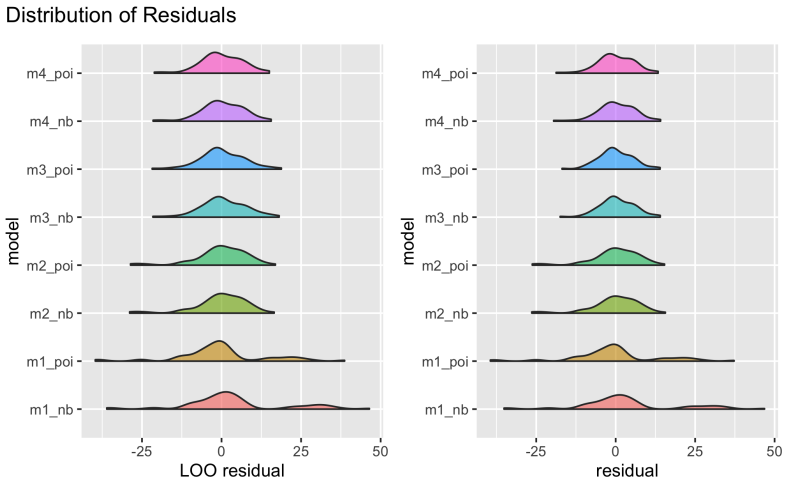


Figure 26. The distributions of residuals (standard and LOO) between the observed mishap counts and mean posterior predictive mishap count. In the LOO residual plot, the spread in the distribution narrows as we progress from M1 to M4.

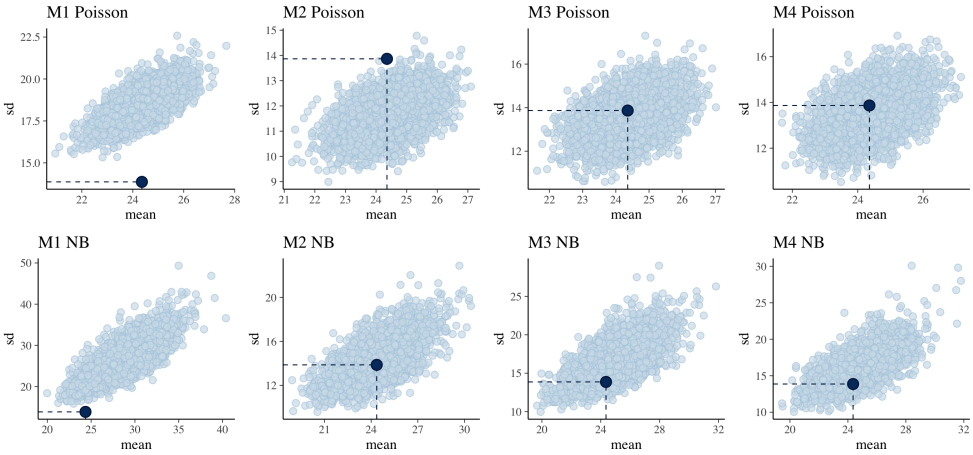


Figure 27. The mean and standard deviation of the posterior predictive distribution, compared with the mean and standard deviation of the observed HCM rate (dark blue point).

F-22, F-15E, and A-10. Among the pilot attribute predictors, we find that x_{lag}, x_{dg}, x_{edu}, and x_{ip} are meaningfully associated with a decrease in HCM rate. Specifically, the 90% credible intervals (CIs) for x_{lag}, x_{dg}, x_{edu}, and x_{ip} exclude 0.0, so we conclude that there is less than a 0.10 probability that the coefficient is nonnegative.

To interpret the magnitude of the effect of personnel factors on HCM rate, recall that the mean HCM count and rate are modeled as

$$\begin{aligned}
\mathbf{E}\left[y_i\right] &= e^{\ln h_i + \beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j}} \\
\mathbf{E}\left[\frac{y_i}{h_i}\right] &= e^{\beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j}} \\
&= e^{\beta_0} \prod_{j=1}^{p} e^{\beta_j x_{i,j}}.
\end{aligned}$$

Therefore, a change in $x_{i,j}$ results in multiplicative effect of $e^{\beta_j x_{i,j}}$ on the expected mishap rate. Using the posterior medians as the point estimates for β_j, the point estimates for β_{lag}, β_{IP}, β_{DG}, and β_{edu} are -0.15, -0.19, -0.27, and -0.15, respectively. The predictors are standardized prior to model fitting so a one unit change in $x_{i,j}$ represents a one standard deviation change in the predictor. In turn, the coefficient estimates imply that a 0.1 standard deviation increase in flight hours in the past year, IP rate, DG rate, and advanced education rate is associated with a 1.4, 1.8, 2.4, and

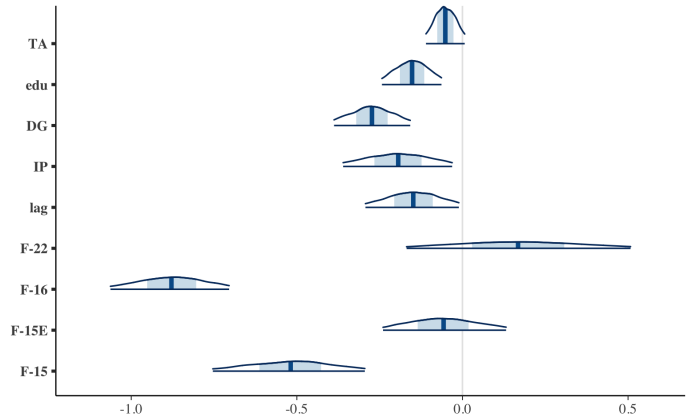


Figure 28. The posterior distribution of model M4 Poisson.

1.4 percent decrease in HCM rate. Using the standard deviations of the predictors reported in Table 5, an equivalent interpretation of the coefficient estimates is that a 10 hour increase in the average flight hours flown in the past year is associated with a 5% decrease in HCM rate. Likewise, a 1% increase in average IP, DG, and advanced education rate is respectively associated with a 1.6%, 4.0%, and 0.9% decrease in HCM rate. For context, from 2008 to 2020 there were on average 120 flight-related HCMs per year with a median cost of $125k and average cost of $2.0M. Assuming total flying hours remain relatively consistent in future years, a 1% decrease in overall HCM rate would result in approximately 1 fewer HCM mishaps per year.

Because the F-35 has recently joined the Air Force fighter inventory and has a nonstandard pilot selection pipeline, we initially removed it from the training set for models M1-M4. In model M5, we include the F-35 data. The posterior distributions are directionally similar to the results from M4, without the F-35, although the magnitudes of the effect sizes are reduced. Since the observational distribution of mishap rates is altered once we include F-35 data, we cannot compare the ELPD_{loo} from model M5 directly with models M1-M4. Given the F-35's unique pilot selection process prior to 2018, the results of M4 are likely more representative of general Air Force trends. Thus, we do not present the results of M5 in detail.

The results from this research are significant from both methodological and policy perspectives. Methodologically, we demonstrate the utility of a Bayesian framework to analyze aviation mishap rates. A Bayesian modeling approach enables the use of predictive projection for feature selection, which has been shown to perform well in comparison to alternative methods of feature selection [67, 60]. Additionally, the Bayesian framework allows for more intuitive interpretations of uncertainty estimates. For example, the 90% credible intervals shown in Figure 28 imply there is a 0.90 probability that the true parameter value lies within the depicted interval. Finally,

although we use default priors in our analysis thus far, the Bayesian approach allows for priors that are motivated by qualitative safety studies and, thus, can incorporate existing qualitative research into the model. We expand on the use of informative priors in the following secion.

Next, our results bolster key findings and recommendations of the NCMAS report. Specifically, the commission proposes that a reduction in flight hours for early career pilots has a causal effect on aviation safety risks. Based on our extensive review of the literature, we contend this research provides some of the first empirical evidence linking pilot flight hours with reductions in HCMs for fighter aircraft. After compiling the weighted average flight hours in the past year for each MDS pilot community, we find that an increase of 10 hours is associated with a 5% decrease in HCM rate. The NCMAS noted that performing such analysis would require improvements in data collection to reduce risk, prevent mishaps, and optimize human performance. Through our detailed compilation of personnel and mishap data at the MDS level, we were able to circumvent many of these challenges and quantify attributes of MDS-specific pilot communities. However, there remain opportunities to improve data collection at the unit level to support future analyses.

Although we are not focused on identifying causal relationships between pilot attributes and mishap rates, associations may be evidence of causal relationships, and non-intuitive associative findings can be useful in highlighting which causal relationships warrant further investigation. We highlight several findings of particular interest which may have policy implications.

First, our results suggest that higher rates of pilots with advanced academic degrees are predictive of lower mishap rates and TA usage has no association with HCM rates. Prior to 2014, many pilots pursued advanced academic degrees primarily to improve their chances of promotion. Beginning in 2014, promotion boards were no

longer allowed to consider advanced academic degrees, and the personnel data reveals a noticeable decline in degree completion following the policy change (Figure 23). The motivation for the policy change was to ensure Air Force officers had enough time to focus on their job performance and to "have a life away from work" ([70]) since most degrees were completed while simultaneously maintaining a full flying schedule. The NCMAS report echoed this line of thought and concluded that pilot fatigue due to excessive non-flying related duties is a causal factor for increased safety risk. Given that our results find a beneficial association between advanced degrees and mishap rates, the previously assumed relationship between advanced degrees and mishap rate warrants reconsideration by the Air Force.

We also find that an increase in the proportion of fighter pilots who are DGs is associated with a decrease in HCM rate. While further analysis is necessary to determine whether the relationship between DG rates and mishap rates is causal, DG status is unique in that it is determined years before an officer learns to fly a fighter aircraft. Thus, we hypothesize that DG status is less likely to be confounded with other pilot attributes such as flying experience. If the relationship between DG rates and mishap rates is causal, the Air Force could recruit commissioning DGs to become fighters pilots via financial incentives or unique career development opportunities such as opportunities to pursue advanced degrees in-residence.

The results presented thus far in this chapter have been submitted to *Safety Science* for review and publication.

3.6 Leveraging Prior Knowledge

Thus far, we have not incorporated subject matter knowledge into our analysis. Although our efforts to use quantitative methods to analyze the relationship between pilot attributes and mishap rates represents a novel contribution, there have been

extensive qualitative research efforts to analyze the causal factors that affect mishap rates. In the NCMAS report, the authors conduct an extensive review of the existing qualitative research and also conduct thousands of interviews in effort to determine potential causes of heightened safety risk. One of the key findings from the NCMAS report is that a reduction in flying time is likely to lead to increased safety risks. The flexibility of Bayesian priors provides a method for incorporating qualitative findings in rigorous quantitative models.

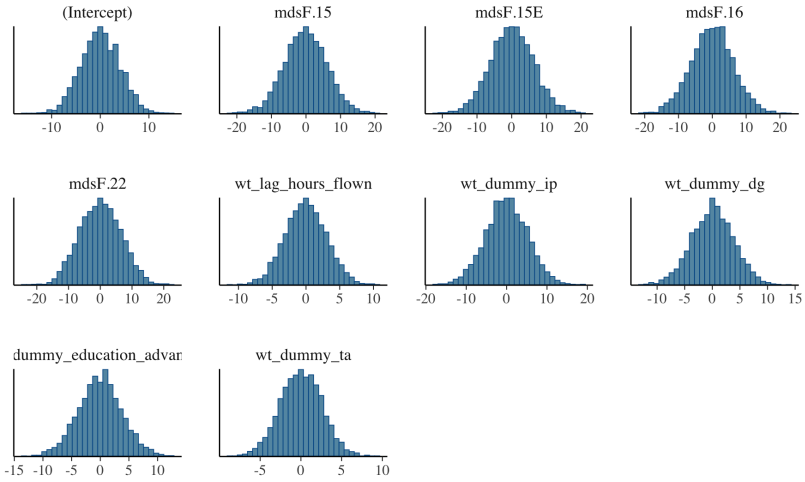


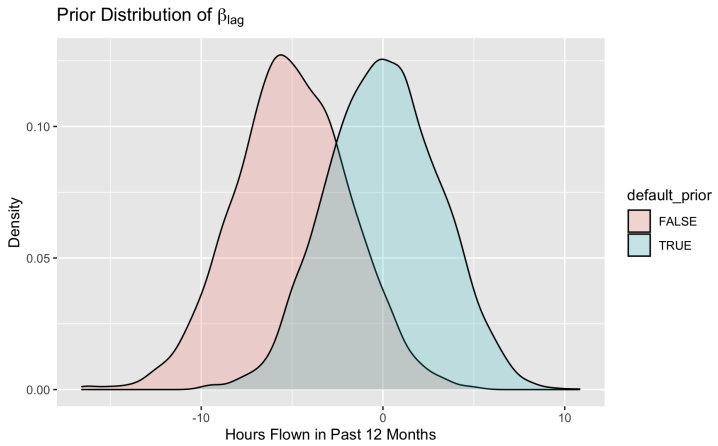
Figure 29. The models presented in section 3.4.2 are fit using noninformative priors on the parameters.

The models presented in section 3.4.2 are fit using non-formative Bayesian priors. Figure 29 shows the non-informative priors that were used to fit models M1-M4. The prior distributions encode our belief about the likely values that each parameter could take on. The non-informative priors are centered at zero to imply the it is unknown if each of the parameter variables has a positive or negative association with mishap rates. We use a Gaussian distribution to represent the belief that extremely high or low values of the parameter are exceedingly unlikely. If we consider the prior on x_{lag}, the range of plausible values are approximately -5 to 5. Adjusting for the standardization of the variables, this implies that a one standard change in recent

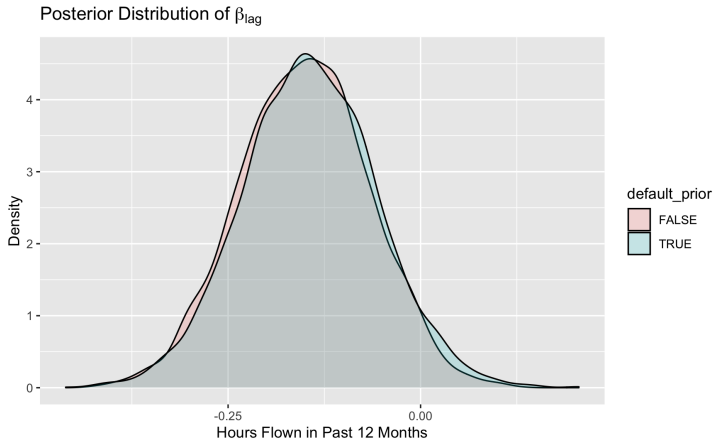
flight hours, ±26.9 hrs, could lead to a nearly 99% reduction or 50% increase in mishap rate. Importantly, the prior on x_{lag} implies that we believe an increase in additional flight hours in last year is equally likely to have a positive or negative effect on mishap rates. If we wish to verify whether there is quantitative evidence of the prior qualitative findings on the relationship between flight hours and safety, then it may be reasonable to use a non-informative prior. However, if we wish to incorporate the findings of prior qualitative studies, we could adjust the prior on x_{lag} to reflect our belief that increase flight hours are likely to reduce mishap rates. Figure 30a depicts an alternative prior for x_{lag} that is centered at -5. If we refit the M4 Poisson model using this informative prior, we find that the posterior distribution of x_{lag}, after conditioning on the available data, is nearly identical to the posterior distribution that was previously generated using a non-informative prior. Thus, in this particular example, the use of an informative prior has no effect on the modeling results. In general, however, Bayesian priors provide a powerful method for incorporating prior qualitative results and subject matter knowledge into a quantitative models.

3.7 Conclusion

In this study, we first present safety data and trends for class A, B, and C flight mishaps from 2007-2020. Notably, we find evidence of abnormalities in the distribution of mishap cost estimates near the threshold between class B and C mishaps. Thus, we focus on all class A, B, and C mishaps, or HCMs. Next, we quantify pilot attributes and present trends in MDS pilot communities using over 15 years of data. Our analysis reveals trends in the pilot communities of each MDS and suggests there are notable differences between pilot communities. We use these data to model HCM rates as a function of pilot attributes using Bayesian regression and conduct feature selection via predictive projection. By comparing both models with



(a) Comparison of a non-informative (default) and informative prior on the parameter, β_{lag}.



(b) Comparison of the posterior distribution of the parameter, β_{lag}, with a default non-informative prior and informative prior

Figure 30. We consider the effect of using an informative prior for the effect of recent flying hours on mishap rates. Despite using an informative prior, the resulting posterior distribution is largely unaffected.

and without personnel factors, we find evidence of a meaningful relationship between personnel factors and HCM rate. Specifically, we find that MDS pilot communities with higher average flight hours in the last year are associated with lower mishap rates. Additionally, we find evidence that higher percentages of pilots who are DGs, are IPs, and have advanced academic degrees are associated with reduced HCM rates. Our efforts to analyze mishap cost estimation behavior, quantify pilot attributes, and model the relationship between pilot attributes and mishap rates represent an applied contribution to the existing literature. Furthermore, our use of Bayesian regression with predictive projection for feature selection represents a valuable methodological contribution. Lastly, we demonstrate the use of Bayesian priors for incorporating the findings of prior qualitative research.

IV. Contribution 3: Using Causal Inference to Analyze Air Force Personnel Evaluation Processes

4.1 Overview and Motivation

In recent years, data science, machine learning, and artificial intelligence methods have been used to solve increasingly complex problems across a wide range of industries. These fields have also become a focal point of the USAF and DoD [71]. However, the majority of machine learning methods and applications focus on predictive modeling and struggle to reveal causal relationships. While prediction is useful for many applications, making policy decisions for the purpose of inducing a desired outcome often requires causal inference.

To motivate the need for causal inference and illustrate the shortfalls of predictive modeling, we revisit the findings presented in Chapter 3. Using a Bayesian regression model, we find that a having higher percentage of pilots who have completed an advanced academic degree is predictive of a lower mishap rate. If this relationship is causal, then Air Force leaders could require pilots to earn advanced degrees in an effort to reduce mishap rate. However, it is possible that high rates of advanced academic degrees are indicative of pilot communities that had a lower operations tempo which afforded them the off-duty time required to pursue a degree. In this hypothetical example, lower operations tempo may be the true cause for the reduced mishap rates, and implementing a policy that requires advanced degrees could cause an increase in mishap rates. Another finding from Chapter III was that a higher average number of flight hours in the past 12 months is associated with a lower mishap rate. Although an Air Force leader might be more inclined to interpret the association between flight hours and mishap rate as a causal relationship based on personal experience, the model does not provide any evidence that flight hours are more likely than advanced

degrees to have a causal effect on mishap rates. Ultimately, the associations revealed by the predictive model in Chapter III cannot be used to develop policies aimed at reducing mishap rates because correlation is not causation.

Given that there are nearly 330,000 active duty members in the Air Force, formal evaluation processes are used throughout the Air Force for a wide range of purposes including selecting distinguished graduates (DG) from training programs, selecting candidates for professional opportunities, and awards. In general, the objective of these evaluation processes is to incentivize and reward attributes or achievements that the Air Force believes are desirable. The results of the evaluation process then affect outcomes of interest such as retention rate, promotion rate, or job performance. The data generated from these processes are then commonly used in regression analyses to determine the relationship between various factors and a performance outcome [72, 73, 74, 75, 76, 77]. However, unlike in many applications, the causal relationships pertaining to evaluation processes are known since they are often dictated by policy. For example, Air Force policy dictates what factors can and cannot be considered for promotions or awards. If knowledge of the evaluation process is not incorporated into the design of the regression model, it is unclear whether the associations revealed by the regression are causal or spurious. Motivated by this uncertainty, we use structural causal models to incorporate knowledge of the evaluation process into our analysis. For readers that are familiar with causal inference, our work is closely related to other efforts to understand "good" and "bad" controls [78].

We show that by leveraging knowledge of the causal structure, we can select a set of explanatory variables such that the resulting regression coefficients do indeed estimate the causal effect of interest. Notably, we derive and use a unique formula for computing the regression coefficients of a multiple linear regression using only the pair-wise covariances of the regression variables. This alternate method for computing

regression coefficients allows us to clearly understand which causal quantities are being estimated by the regression coefficients. We show that, depending on which causal relationship we are trying to estimate, the causal structure dictates which predictors or covariates must be controlled for to correctly identify the causal effect of interest. Furthermore, we show that adding additional predictors is not always desirable because it can either lead to biased estimates or inefficient estimation.

In this chapter, we first provide further background on evaluation processes and motivate the need for causal inference. Next, we provide a brief introduction to causal inference with a narrow focus on the topics necessary for understanding the following sections. We then present a unique method for computing regression coefficients from pair-wise covariances of the predcitors. Lastly, we present four different hypothetical causal models that represent the types of evaluation processes that may be common in the Air Force. For each of the four models, we show that a regression analysis with different sets of predictors lead to estimates of different causal quantities.

4.2 Related Works

There are numerous examples of research involving the analysis of data involving evaluation processes. First, several studies have investigated the connection between attributes of pilot training candidates and their performance during undergraduate pilot training (UPT) [72, 73, 74]. Since UPT candidates must go through a selective process, the performance of candidates in UPT is dependent on the UPT candidate evaluation process. Next, Keller et. al. [77] study the Air Force's enlisted promotion system. Specifically, they train a model that predicts the probability of promotion to master sergeant using the factors that are considered in the Weighted Airman Promotion System (WAPS). WAPS is an evaluation process that has a causal effect on the future promotion of enlisted members to master sergeant. Analyzing the promotion

rate to master sergeant should account for the effect of the evaluation process on the data-generating process. Lastly, King et. al. [79] study the retention rate of women in the military at a fixed number of years into their career. Awards, promotions, and availability of career opportunities are all subject to evaluation processes, and are likely to affect an airman's decision to stay in the Air Force. Each of these studies fails to acknowledge the causal structure of the evaluation processes that generated the data. Therefore, it is unclear whether the estimated associations are causal or spurious.

4.3 Causal Inference Theory

4.3.1 Structural Causal Models

We provide a brief introduction to causal inference with a narrow focus on the topics relevant to the work presented in future sections. This introduction is heavily inspired by the introduction in the appendix of Cinneli et. al. [78]. Pearl et. al [80] also provide an in-depth introduction to linear causal models.

Causal inference often involves estimating causal relationships between variables using an observational data set. This is notably different than many applications of machine learning and data science where the goal is to build a predictive model that relies on identifying association, not causation. Structural Causal Models (SCM) are central to causal inference and provide a framework to understand the differences between predictive and causal inference. In predictive analytics, there are no assumptions about the causal relationships that led to the observed data. In causal inference, we assume there is an underlying causal structural and the causal structure can be expressed via an SCM. An SCM consists of three components; it consist of a set of endogenous variables, exogenous variables, and a set of functions. Ultimately, an SCM describes a data generating process (DGP) that induces a joint probability dis-

tribution over the set of endogenous variables. We use a concrete example to further introduce concepts related to SCMs. An SCM, M, is defined as follows,

$$M = \begin{cases} Z \leftarrow f(U_z) & = U_z \\ X \leftarrow f(Z, U_x) & = \lambda_{zx} Z + U_x \\ Y \leftarrow f(X, Z, U_y) & = \lambda_{zy} Z + \lambda_{xy} X + U_y \\ U_x, U_z, U_y & \sim P(u_x, u_z, u_y) \end{cases} \tag{3}$$

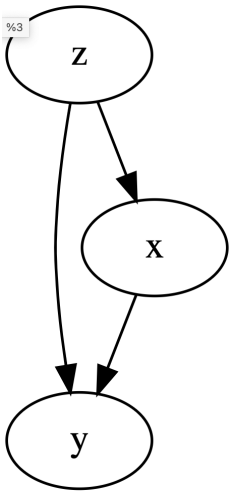


Figure 31. The SCM M induces a DAG. The nodes represent the endogenous variables. The directed edges represent causal relationships.

where the set of variables $\mathbf{V} = \{X, Z, Y\}$ are endogenous variables; the set of variables $\mathbf{U} = \{U_x, U_z, U_y\}$ are exogenous variables; and the functions defining each endogenous variable define the causal relationships in the model. The model M then induces a joint distribution over the endogenous variables. An observational dataset is a sample from the induced joint distribution. Once an SCM is defined, a corresponding causal graph can be drawn. The corresponding graph is a directed graph that consists of a node for each endogenous variables and an edge for each causal relationship. Figure 31 displays the DAG defined by M. We only consider

SCMs that have a corresponding graph that is a directed acyclic graph (DAG). That is, causal relationships must be one-directional so that causal effects only flow from parent node to child node[1]. The causal effects in M are represented by $\lambda_{z,x}, \lambda_{z,y}$, and $\lambda_{z,y}$. In our work, we restrict our focus to linear causal models. In general, causal effects need not be linear.

Once an SCM is defined, it is possible to model the effect of an intervention in a straightforward manner. An intervention involves setting an endogenous variable to a particular value or, equivalently, replacing $X \leftarrow f(Z, U_x)$ with $X \leftarrow x$. We denote an intervention using $do(X = x)$ and the modified SCM as M_x. The modified SCM can be represented with a modified causal graph where all incoming edges to X are removed. This then induces a new joint distribution called the interventional distribution which is denoted as $P(\mathbf{V}|do(X = x))$. The interventional distribution is unobserved, but can be estimated using the observational distribution if the conditional exchangeability or unconfounded assumption is met. The interventional distribution can then be used to determine the average causal effect (ACE), $E[Y \mid do(x + 1)] - E[Y \mid do(x)]$. In M_x, the ACE on Y of an intervention on X is $\lambda_{x,y}$. A key insight is that, if we can estimate the interventional distribution, then we are able to model counterfactuals. However, progressing from an observational distribution to an interventional distribution is only made possible by knowledge of the underlying SCM.

In our work, we use SCMs to encode our knowledge of Air Force evaluation processes. For example, based on Air Force policy we know what factors have a causal effect on the Weighted Airman Promotion System (WAPS) score. The WAPS score then has a causal effect on whether an enlisted airman is promoted. These causal relationships can be captured in an SCM and illustrated via a DAG. In the next section,

[1]There are limited works in the area of cyclic causal structures [81, 82, 83], but the vast majority of existing literature is related to the study of SCMs that define a DAG.

we explore how path analysis of a DAG can aid in understanding which variables to include in a regression model.

4.3.2 Flow of Causation and Association in a DAG

Linear regression provides a means for estimating the association between variables. In a regression with Y as the response and X as the lone predictor, the regression coefficient is given by,

$$\begin{aligned} \beta_{yx} &= \frac{\text{Cov}(Y, X)}{\text{Var}(X)} \\ &= \frac{\sigma_{xy}}{\sigma_x^2} \\ &= \rho_{xy}\frac{\sigma_y}{\sigma_x} \end{aligned} \tag{4}$$

where $\rho_{xy} = \rho_{yx} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$ is the correlation between X and Y. Without loss of generality, if X and Y are standardized variables with a variance of one, then the regression coefficient is equivalent to the correlation between X and Y. If we assume the variables in M are standardized with variance of 1, we will later show that we can use path analysis of the DAG to compute the regression coefficient as $\beta_{yx} = \rho_{xy} = \sigma_{xy} = \lambda_{xy} + \lambda_{zx}\lambda_{zy}$. Notably, the regression coefficient does not equal the known causal effect, λ_{xy}. However, if we fit a regression with both X and Z as predictors, the regression coefficient on X, $\beta_{yz.z} = \lambda_{xy}$. The variable Z is, therefore, a "good" control because it helps identify the causal effect once added to the regression.

In general, we can determine the flow of non-causal and causal association in any DAG by analyzing the types of paths that connect each pair of variables. Furthermore, an analysis of the paths in a DAG reveals which variables must be included and excluded to ensure that correlation equals causation. There are three possible types of paths in a DAG:

1. Chain: A chain is a path of the form $A \rightarrow B \rightarrow C$. Association flows bi-directionally along a chain. Conditioning on B blocks the flow of association.

2. Fork: A fork is a path of the form $A \leftarrow B \rightarrow C$. Association flows bi-directionally along a fork. Conditioning on B blocks of the flow of association.

3. Collider: A collider is a patch of the form $A \rightarrow B \leftarrow C$. Association does not flow along a collider. Conditioning on B opens the flow of association.

Association flows along a path between two nodes if there is not a collider somewhere along the path. If we condition on the middle node in a chain or fork, the flow of association along the path is blocked. The collider path is unique in that it blocks the flow of association by default. Conditioning on the middle node of a collider then opens the flow of association. The collider is important because it means that adding more variables to a regression can be harmful. A set of variables, $\mathbf{Z}$, is called a valid adjustment set for identifying the causal effect of X on Y if it blocks the flow of all non-causal association. A valid adjustment set can often[2] be found using the following process [84]:

1. List all paths between X and Y

2. If a path consists entirely of chains, then the path is a direct causal path and should not be blocked. All other paths are indirect paths.

3. If an indirect path has a collider, then the path is blocked by default.

4. If an indirect path has a fork and no colliders, add the confounding variable (i.e. middle node) in the fork to $\mathbf{Z}$.

[2]There are several unique DAGs where this heuristic may be insufficient for constructing a valid adjustment set. However, that is beyond the scope of our work and is not pertinent to the DAGs we consider.

Assuming we have a valid adjustment set, $\mathbf{Z}$, for identifying the causal effect of X on Y, we can estimate the causal effect of X on Y using a regression with Y as a dependent variable, X as an independent variable, and the set of variables, $\mathbf{Z}$, included as additional predictors or controls in the regression. The resulting fitted coefficient on X is denoted as $\beta_{yx.\mathbf{Z}}$ to indicate that it is from a regression where X is regressed on Y while controlling for the variables in $\mathbf{Z}$. Then $\beta_{yx.\mathbf{Z}}$ will be an unbiased estimate of the total causal association of X on Y. Thus, controlling for a valid adjustment set in a regression model allows us to interpret correlation as causation.

4.3.3 Analytic Results for Linear Causal Models

In linear SCMs, each edge represents a causal effect that is captured by a single parameter λ_{xy} (or l_{xy}) where X is the parent node and Y is the child node. For any two nodes in a DAG, X and Y, the covariance between X and Y can be computed using the unblocked paths (i.e. paths without a collider) from X to Y. For each unblocked path, we compute flow of association along the path by taking the product of the linear coefficients of each edge along the path, as well as the variance of the source node of each path. We then sum the product terms for each path to determine the covariance, σ_{xy}. In the example M, we compute the covariance between X and Y as $\sigma_{xy} = \sigma_x^2\lambda_{xy} + \sigma_z^2\lambda_{zx}\lambda_{zy}$. Thus, the covariance between any pair of variables (nodes) can be written as a function of the causal edge coefficients and variance of the source node. For any subset of the endogenous variables, the covariance matrix for p variables, $X_1, X_2, ..., X_p$, is defined as,

$$\text{Cov}\left(\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix}\right) = C_x = \begin{bmatrix} \sigma^2_{X_1} & \sigma_{X_1X_2} & \cdots & \sigma_{X_1X_p} \\ \sigma_{X_1X_2} & \sigma^2_{X_2} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \\ \sigma_{X_1X_p} & \cdots & & \sigma^2_{X_p} \end{bmatrix} \tag{5}$$

Next, suppose we select p predictors, $X_1, X_2, \ldots, X_p$, from the set of endogenous variables in an SCM. Additionally, we select one endogenous variable to be the response variable, Y. Let $\mathbf{X}$ be a $n \times (p+1)$ data matrix. The n row vectors represent observations and each of the columns vectors, $\vec{1}, \vec{X_1}, \vec{X_1}, \ldots, \vec{X_p}$, represent the observed values for each of the p predictors along with one column for the intercept. Let $\vec{Y}$ be a $n \times 1$ vector consisting of observations of the response variable,

$$\begin{aligned} \vec{Y} &= \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \\ \mathbf{X} &= \begin{bmatrix} | & | & & | \\ \vec{1} & \vec{X_1} & \cdots & \vec{X_p} \\ | & | & & | \end{bmatrix} \end{aligned} \tag{6}$$

Using the normal equations, $\mathbf{X'X}\hat{\beta} = \mathbf{X'}\vec{Y}$, the least squares regression coefficients are given by $\hat{\beta} = \mathbf{X'X}^{-1}\mathbf{X'}\vec{Y}$. However, computing the regression coefficients in this manner makes it difficult to understand the connection between the regression coefficients and the causal edge coefficients. In contrast, if the regression coefficients can be computed using the covariance matrix of the predictors, C_x, then the regression coefficients can be expressed as a function of the causal edge coefficients. Thus, we derive a formula for computing $\hat{\beta}$ using the covariance matrix, $\mathbf{C}_x$. This derivation

of the regression coefficients is partially presented in an online forum [85] but, to the best of our knowledge, has not been published. Furthermore, to the best of our knowledge, it has has not been used to analyze linear causal models.

Let $\mathbf{C}_x$ be the $p \times p$ covariance matrix of the vector of predictor variables in the regression. Let $\vec{C}_{xy}$ be the vector of covariances between each predictor, X_i, and the response variable Y. We will show that $\hat{\beta} = C_x^{-1}\vec{C}'_{xy}$.

We first define components of the normal equations $\mathbf{X'X}\hat{\beta} = \mathbf{X'}\vec{Y}$.

$$\begin{aligned}
\mathbf{X'X} &= \begin{bmatrix} — & \vec{1}' & — \\ — & \vec{X_1}' & — \\ & \vdots & \\ — & \vec{X_p}' & — \end{bmatrix} \begin{bmatrix} | & | & & | \\ \vec{1} & \vec{X_1} & \cdots & \vec{X_p} \\ | & | & & | \end{bmatrix} \\
&= \begin{bmatrix} n & \vec{1}'\vec{X_1} & \cdots & \vec{1}'\vec{x_p} \\ \vec{X_1}'\vec{1} & \vec{X_1}'\vec{X_1} & \cdots & \vec{X_1}'\vec{X_p} \\ \vdots & \vdots & \ddots & \vdots \\ \vec{X_p}'\vec{1} & \vec{X_p}'\vec{X_1} & \cdots & \vec{X_p}'\vec{X_p} \end{bmatrix} \\
\mathbf{X'}\vec{Y} &= \begin{bmatrix} \vec{1}'\vec{Y} \\ \vec{X_1}'\vec{Y} \\ \vdots \\ \vec{X_p}'\vec{Y} \end{bmatrix}
\end{aligned} \tag{7}$$

The regression coefficients, $\hat{\beta}$, are the solution to the linear system of equations, $\mathbf{X'X}\hat{\beta} = \mathbf{X'}\vec{Y}$. We define the augmented matrix, A,

$$A = \left[\begin{array}{cccc|c} n & \vec{1}'\vec{X_1} & \cdots & \vec{1}'\vec{x_p} & \vec{1}'\vec{Y} \\ \vec{X_1}'\vec{1} & \vec{X_1}'\vec{X_1} & \cdots & \vec{X_1}'\vec{X_p} & \vec{X_1}'\vec{Y} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \vec{X_p}'\vec{1} & \vec{X_p}'\vec{X_1} & \cdots & \vec{X_p}'\vec{X_p} & \vec{X_p}'\vec{Y} \end{array}\right]. \tag{8}$$

We then apply Gaussian elimination to the first column of A. First, we use the row operation $\frac{1}{n}R_1 \to R_1$ such that $A_{11} = 1$. First, we multiply each row, R_i, by $(1/n)$ such that $\frac{1}{n}R_i \to R_i$.

$$A = \left[\begin{array}{cccc|c} 1 & \overline{X}_1 & \cdots & \overline{X}_p & \overline{Y} \\ 0 & \frac{\vec{X_1}'\vec{X_1}}{n} - (\overline{X}_1)^2 & \cdots & \frac{\vec{X_1}'\vec{X_p}}{n} - \overline{X}_1\overline{X}_p & \frac{\vec{X_1}'\vec{Y}}{n} - \overline{X}_1\overline{Y} \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & \frac{\vec{X_p}'\vec{X_1}}{n} - \overline{X}_p\overline{X}_1 & \cdots & \frac{\vec{X_p}'\vec{X_p}}{n} - (\overline{X}_p)^2 & \frac{\vec{X_p}'\vec{Y}}{n} - \overline{X}_p\overline{Y} \end{array}\right]. \tag{9}$$

Notice that $A_{i+1,j+1} = \frac{\vec{X_i}'\vec{X_j}}{n} - \overline{X}_i\overline{X}_j$. Recall that the covariance between two variables, X and Y, is defined as $\text{Cov}(X,Y) = \mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y]$. Therefore, if we exclude the intercept term, β_0, the augmented matrix implies that, asymptotically, the remaining regression coefficients solve $C_x\hat{\beta} = \vec{C}'_{xy}$. If the covariance matrix, C_x, is invertible, $\hat{\beta} = C_x^{-1}\vec{C}'_{xy}$. Assuming we have full knowledge of the SCM, this allows us to express the regression coefficients as functions of the known causal edge coefficients.

To demonstrate the importance of this result, consider the DAG shown in Figure 32. Suppose we wish to fit a regression with Y as the response variable and the set $\{X_1, X_2, S\}$ as controls. We first compute the covariance matrix for the control variables,

$$C_x = \begin{bmatrix} v_{x1} & 0 & l_{x1s}v_{x1} \\ 0 & v_{x2} & l_{x2s}v_{x2} \\ l_{x1s}v_{x1} & l_{x2s}v_{x2} & v_s \end{bmatrix} \tag{10}$$

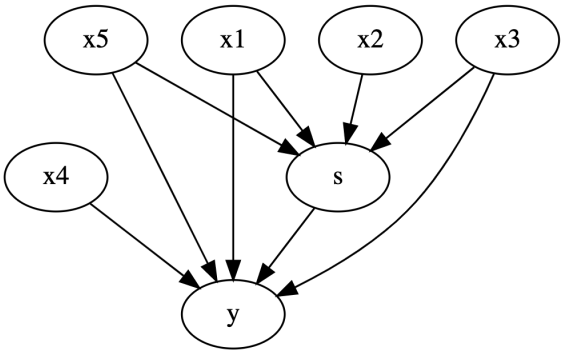


Figure 32. In this hypothetical model, a set of attributes, $X_1, X_2, ..., X_5$, describe an individual. The evaluation process uses a subset of these attributes to generate a score, S. A subset of the attributes also have an effect on the performance outcome, Y. Lastly, the evaluation score itself has an effect on Y.

and the covariance vector,

$$\vec{C}'_{xy} = \begin{bmatrix} l_{sy}l_{x1s}v_{x1} + l_{x1y}v_{x1} \\ l_{sy}l_{x2s}v_{x2} \\ l_{sy}v_s + l_{x1s}l_{x1y}v_{x1} + l_{x3s}l_{x3y}v_{x3} + l_{x5s}l_{x5y}v_{x5} \end{bmatrix} \tag{11}$$

where l_{ij} is edge coefficient from node i to node j and v_i is the variance of node i. Then we compute the regression coefficients as

$$\begin{bmatrix} \beta_{x_1} \\ \beta_{x_2} \\ \beta_s \end{bmatrix} = C_x^{-1}\vec{C}'_{xy} = \begin{bmatrix} \frac{l_{x1s}^2 l_{x1y}v_{x1} + l_{x1s}l_{x3s}l_{x3y}v_{x3} + l_{x1s}l_{x5s}l_{x5y}v_{x5} + l_{x1y}l_{x2s}^2 v_{x2} - l_{x1y}v_s}{l_{x1s}^2 v_{x1} + l_{x2s}^2 v_{x2} - v_s} \\ \frac{l_{x2s}(l_{x3s}l_{x3y}v_{x3} + l_{x5s}l_{x5y}v_{x5})}{l_{x1s}^2 v_{x1} + l_{x2s}^2 v_{x2} - v_s} \\ \frac{l_{sy}l_{x1s}^2 v_{x1} + l_{sy}l_{x2s}^2 v_{x2} - l_{sy}v_s - l_{x3s}l_{x3y}v_{x3} - l_{x5s}l_{x5y}v_{x5}}{l_{x1s}^2 v_{x1} + l_{x2s}^2 v_{x2} - v_s} \end{bmatrix}. \tag{12}$$

Despite the relatively small DAG and only three controls, the relationship between the regression coefficients and the causal edge coefficients is non-trivial.

Other analyses of linear causal models use a recursive method to compute the regression coefficients via the covariances between variables [80]. The recursive method is derived by Cramer[3] [86]. In the case where there are only $p = 2$ predictors, the

[3]The full recursive method as derived by Cramer, where the variables are simply numbered

recursive method for computing the regression coefficient reduces to,

$$\beta_{yx.z} = \frac{\sigma_{yx} - \sigma_{xz}\sigma_{yz}}{1 - \sigma_{xz}^2}$$

where $\sigma_{i,j}$ is the covariance between two variables. For models with more than two predictor variables, $p > 2$, the recursive method becomes difficult to use. Furthermore, the recursive method only allows for computing one regression coefficient at a time and it must be applied repeatedly for each separate predictor. Thus, using $\hat{\beta} = C_x^{-1}\vec{C}'_{xy}$ allows us to more easily analyze regression with more than two predictors.

4.4 Air Force Evaluation Processes

We assume that there exists a performance metric, Y, that the Air Force seeks to maximize. For example, Y could be metrics such as performance during pilot training, retention rate, or commanders' peer ratings. Next, we assume every member of the Air Force can be described by a finite set of attributes, $X_1, X_2, ..., X_p$. Only a subset of these attributes have a direct causal effect on the performance metric, Y. The goal of an Air Force evaluation process, then, is to identify individuals who have the attributes necessary to improve the future performance or capabilities of the Air Force. When instituting an evaluation process, the Air Force must decide which of the p attributes will be considered and how they will be weighted. Ultimately, we assume that an evaluation process results in a score, S. Furthermore, we assume that the evaluation score then has an effect on Y since individuals who rated highly in

$1, 2, ..., n$, is given by the equations,

$$\begin{aligned} \rho_{12.34...n} &= \frac{\rho_{12.34...n-1} - \rho_{1n.34...n-1}\rho_{2n.34...n-1}}{\sqrt{\left(1 - \rho_{1n.34...n-1}^2\right)\left(1 - \rho_{2n.34...n-1}^2\right)}} \\ \beta_{12.34...n} &= \rho_{12.34...n}\frac{\sigma_{1.34...n}}{\sigma_{2.34...n}} \end{aligned} \tag{13}$$

an evaluation are likely to have an advantage over their peers. Figure 32 illustrates a hypothetical evaluation process via a DAG. While this characterization of the Air Force selection processes is simplistic, it is sufficient to demonstrate the utility of causal inference.

There are several common characteristics of the DAGs pertaining to evaluation processes which make them particularly interesting. First, the evaluation score, S, will nearly always be the child node to many parent nodes. Each of the parent nodes represents an attribute that has an effect on the evaluation score. Since S is the child node of many attribute nodes, there will be a collider path, $X_i \rightarrow S \leftarrow X_j$, between any two attributes that have an effect on S. These paths are blocked by default due to the presence of a collider. However, including S as a control in a regression will open a path between every pair of attributes used for evaluation. Thus, the decision to include S as a control will necessitate the need to include every other parent of S. Second, it is common for an attribute, X_i, to have a causal effect on S and Y. This implies that the evaluation process is indeed considering factors that affect the performance outcome, Y. If we consider the resulting DAG, there are two causal paths from X_i to Y: $X \rightarrow Y$ and $X \rightarrow S \rightarrow Y$. We refer to the causal path $X \rightarrow S \rightarrow Y$ as the inorganic effect since the path is entirely due to an evaluation process that has been constructed by the Air Force. Typically, blocking a causal path is undesirable because it precludes us from accurately estimating the effect of an intervention, $do(X_i = x)$. However, in the case of evaluation processes, it may be desirable to block the inorganic path to determine whether the attributes considered during the evaluation process have a direct effect on the performance outcome. Lastly, it is natural to be interested in the causal effect of S on Y since this describes the effect an evaluation process has on future outcomes. For example, we may be interested in the causal effect of earning DG status on an Air Force career. Since many of the

attributes that have an effect on S also have an effect on Y, there are numerous backdoor paths between S and Y. Accurate identification of the causal effect of S on Y requires controlling for every attribute that affects both S and Y.

An example of an evaluation process in the Air Force is the process for determining distinguished graduates (DG) from the United States Air Force Academy. Graduates from the Air Force Academy are evaluated on their performance in academics, physical fitness, and military leadership. Each of the three components are scored separately, and then combined to generate an overall performance score. Graduates with a overall performance score in the 10% of their graduating class receive the DG designation. Importantly, the DG designation is a permanent part of an officer's career records and can influence whether an officer is selected for various career opportunities in the future. Since the DG designation becomes a factor in future evaluations, the DG selection process at USAFA can have long-lasting effects on an officer's career. As is the case with most Air Force evaluation processes, we have extensive knowledge of the causal relationships that lead to DG status. We can leverage this knowledge in the form of a structural causal model.

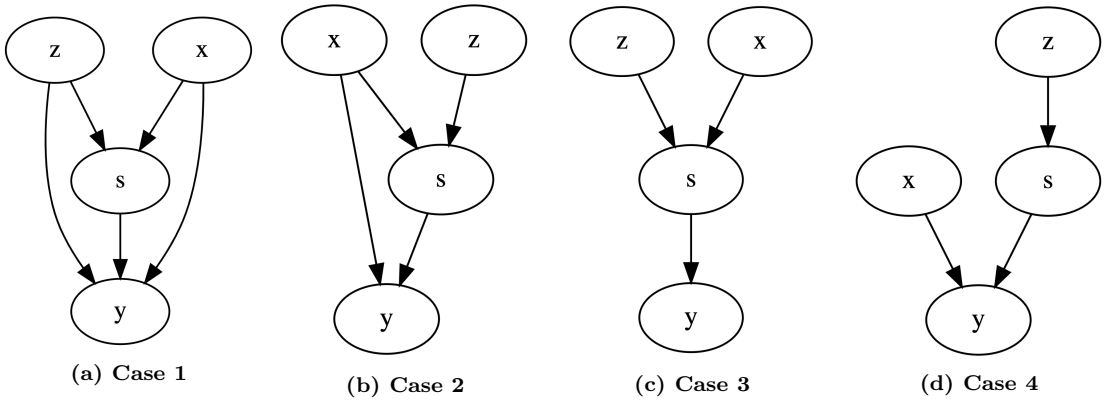
As annother example, we consider the Weighted Airman Promotion System (WAPS). WAPS is used to award points to enlisted airman based on factors such as physical fitness and knowledge test scores. The point total is then used to for promotion decisions to the rank of staff sargeant and technical sergeant. Separately, suppose we wish to do a study on factors that may have a causal effect on retention rates and one factor we wish to consider are knowledge test scores. We may hypothesize that individuals who score high on knowledge tests experience greater job satisfaction. If so, a possible policy change, or intervention, would be to provide increased job training for airman. If we wish to estimate the causal effect of knowledge test score on retention rates via regression, it is unclear whether WAPS scores or other WAPS components

should be included in the regression as controls. We revisit this example in future sections.

Our methodology for analyzing USAF evaluation processes involves the following steps. First, we propose a hypothetical causal model via an SCM. The hypothetical causal models that we propose are meant to inspired by the commonly characteristics in USAF evaluation process. Second, we consider the consequences of fitting a regression with Y as the response and different sets of variables included in the regression. We use $\hat{\beta} = C_x^{-1}\vec{C}'_{xy}$ to express the each of the regression coefficients as a function of the causal edge coefficients in the SCM. Then, we can compare the regression coefficients that result from changing the set of controls included in the regression, and we can compute the bias term from including bad controls.

4.4.1 Case 1

We first consider the case when the set of attributes that are used for evaluation are the same set of attributes that directly effect the performance outcome, Y. Figure 33a depicts the the DAG associated with case 1. We assume there are two attributes, X and Z, that effect the performance outcome, Y, and the evaluation process also considers the same two attributes. The causal relationship between a parent node, i,



(a) Case 1 (b) Case 2 (c) Case 3 (d) Case 4

Figure 33. We consider four different causal models that are representative of evaluation processes in the Air Force.

and child node, j, is represented by l_{ij}. The structural causal model is,

$$M_1 = \begin{cases} Z \leftarrow f(U_z) & = U_z \\ X \leftarrow f(U_x) & = U_x \\ S \leftarrow f(X, Z, U_s) & = l_{zs}Z + l_{xs}X + U_s \\ Y \leftarrow f(X, Z, U_y) & = l_{zy}Z + l_{xy}X + U_y \\ U_x, U_z, U_s, U_y & \sim P(u_x, u_z, U_s, u_y) \end{cases} \tag{14}$$

There are several causal quantities that may be of interest. First, we may be interested estimating the causal effect l_{sy} which can be interpreted as the causal effect of an evaluation score on a performance outcome. For example, the Air Force may be interested in estimating the causal effect of WAPS scores on future career outcomes. A naive approach would be to regress S on Y. This would result in the regression coefficient $\hat{\beta}_{ys} = \frac{l_{sy}v_s + l_{xs}l_{xy}v_x + l_{zs}l_{zy}v_z}{v_s}$, where v_i denotes the variance of the variable or node i. We can clearly see that the regression coefficient does not estimate the causal quantity of interest, l_{sy}. An analysis of the DAG reveals that there are two backdoor paths: $S \leftarrow X \rightarrow Y$ and $Y \leftarrow Z \rightarrow S$. Since these paths are forks, they are unblocked and allow for non-causal association to flow from S to Y. A valid adjustment set to identify the causal effect of S on Y would be Z, X. If we fit a regression with X, Z, and S as predictors, we can use $\hat{\beta} = C_x^{-1}\vec{C}'_{xy}$ to determine that $\beta_{ys.zx} = l_{sy}$. Thus, the regression coefficient can be interpreted as an estimate of the causal effect of S on Y.

Additionally, we find that $\beta_{yx.zs} = l_{xy}$ and $\beta_{yz.sx} = l_{zy}$. The causal effect l_{xy} is the direct effect of X on Y. In the context of an evaluations process, this quantity may be of interest because it is the organic effect that an attribute has on the performance metric. However, this is not the only causal path from X to Y. The path $X \rightarrow S \rightarrow Y$

is also a causal path. We refer to this causal effect as an inorganic effect on Y since the path exists due to an evaluation process that is constructed by the Air Force. In this particular example, it may be desirable to intentionally block a causal path to aid in the estimation of a direct effect.

If we fit a regression that includes X and Z as predictors, we find that $\beta_{yx.z} = l_{sy}l_{xs} + l_{xy}$. This shows that if we exclude the evaluation process score, S, from the regression, the resulting regression coefficients will be a function of the inorganic causal effect of X on Y due to the evaluation process. If X has a large effect on the evaluation score S, the regression coefficient, $\beta_{yx.z}$, may be disproportionately influenced by l_{xs}. In Keller et. al. [77], they use the attributes that are factored into WAPS, without including WAPS score itself in the model, to predict future career performance. This process is analogous to fitting a regression without S as a control. However, Keller's intent was to determine whether attributes factored into the WAPS score are indeed attributes that lead to future success, i.e., is there is an organic, direct affect on future success. By excluding WAPS score from the model, there is the possibility that the estimated association between a WAPS component and future success is unintentionally measuring the inorganic causal effect of the evaluation process.

Lastly, we note that $\beta_{yx} = \beta_{yx.z} = l_{sy}l_{xs} + l_{xy}$. That is, we can estimate the total causal effect of X on Y by either including Z as an additional control or by including no additional controls. However, if we simulate 100 observations from the SCM and fit a regression with and without, Z, we find that the 95% CI for β_{yx} is significantly wider than the 95% CI for $\beta_{yx.z}$. While our focus is on whether regression coefficients estimate the proper causal quantities, it is important to note that the decision to include or exclude additional controls can affect efficiency.

Table 9 provides a summary of the results pertaining to Case 1. The simulated

numeric results in Table 9 are generated using $l_{xy} = 1, l_{zx} = 2, l_{zy} = 3, v_x = 17$, and $v_z = 4$.

4.4.2 Cases 2, 3, and 4

We consider three additional causal models that are representative of evaluations processes. For these cases, we present results without an in-depth discussion of the differences in model specifications. The DAGs for these three models are shown in Figure 33. In Case 2, we consider the scenario where the evaluation process accounts for two variables, X and Z, but only X has a causal effect on Y. In Case 3, we consider the scenario where attributes used for evaluation have no direct effect on Y. Lastly, Case 4 considers the scenario where the evaluation process accounts for Z, but only X has a causal effect on Y. For each of these three cases, Tables 10-12 respectively summarize the the causal quantities estimated by the regression coefficients, depending on the which set of controls or included in the regression.

4.5 Conclusion

The work presented in this chapter provides a framework for estimating causal effects using data that are associated with evaluation processes. SCMs and DAGs provide a mechanism to incorporate existing knowledge of the evaluation processes. If we assume a linear causal model, we can use the DAG to compute the covariance between any two nodes in the DAG as a function of the causal edge coefficients. The pair-wise covariances can then be used to compute the regression coefficients via $\hat{\beta} = C_x^{-1}\vec{C}'_{xy}$. This allows us to express each regression coefficient in terms of the causal edge coefficients in the DAG. Thus, depending on which causal quantity is of interest, we can select the proper controls to ensure that the regression coefficients have a causal interpretation.

Table 9. Analytic and simulation results for a regression with varying set of controls on data generated by Case 1.

Model 1 Analysis			
Regression	Analytic Solution	Expected Numeric	Simulated 95% CI ($n = 100$)
$y \sim s$	$\beta_s = \frac{l_{sy}v_s + l_{xs}l_{xy}v_x + l_{zs}l_{zy}v_z}{v_s}$	0.43	$(0.42, 0.44)$
$y \sim x$	$\beta_x = l_{sy}l_{xs} + l_{xy}$	0.7	$(-0.03, 0.78)$
$y \sim x + z$	$\beta_x = l_{sy}l_{xs} + l_{xy}$	0.7	$(0.68, 0.70)$
	$\beta_z = l_{sy}l_{zs} + l_{zy}$	1.8	$(1.79, 1.81)$
$y \sim x + z + s$	$\beta_x = l_{xy}$	0.10	$(-0.35, 0.42)$
	$\beta_z = l_{zy}$	0.30	$(0.14, 0.52)$
	$\beta_s = l_{sy}$	0.60	$(-0.29, 1.26)$

Table 10. Analytic and simulation results for a regression with varying set of controls on data generated by Case 2.

Model 2 Analysis			
Regression	Analytic Solution	Expected Numeric	Simulated 95% CI ($n = 100$)
$y \sim s$	$\beta_s = l_{sy} + \frac{l_{xs}l_{xy}v_x}{v_s}$	0.31	$(0.31, 0.32)$
$y \sim x$	$\beta_x = l_{sy}l_{xs} + l_{xy}$	0.7	$(0.42, 0.82)$
$y \sim z$	$\beta_x = l_{sy}l_{zs}$	1.2	$(0.97, 1.31)$
$y \sim x + z$	$\beta_x = l_{sy}l_{xs} + l_{xy}$	0.7	$(0.68, 0.70)$
	$\beta_z = l_{sy}l_{zs}$	1.2	$(1.18, 1.21)$
$y \sim x + s$	$\beta_x = l_{xy}$	0.1	$(0.08, 0.10)$
	$\beta_s = l_{sy}$	0.3	$(0.30, 0.30)$
$y \sim x + z + s$	$\beta_x = l_{xy}$	0.1	$(-0.77, 0.24)$
	$\beta_s = l_{sy}$	0.3	$(0.22, 0.73)$
	$\beta_z = 0$	0.0	$(-1.73, 0.30)$

Table 11. Analytic and simulation results for a regression with varying set of controls on data generated by Case 3.

Model 3 Analysis			
Regression	Analytic Solution	Expected Numeric	Simulated 95% CI ($n = 100$)
$y \sim s$	$\beta_s = l_{sy}$	0.3	$(0.30, 0.30)$
$y \sim x$	$\beta_x = l_{xs}l_{sy}$	0.6	$(0.26, 0.73)$
$y \sim z$	$\beta_x = l_{sy}l_{zs}$	1.2	$(1.02, 1.26)$
$y \sim x + s$	$\beta_x = 0.0$ $\beta_s = l_{sy}$	0.0 0.3	$(-0.01, 0.01)$ $(0.30, 0.30)$
$y \sim x + s + z$	$\beta_x = 0.0$ $\beta_s = l_{sy}$ $\beta_z = 0.0$	0.0 0.3 0.0	$(-0.40, 0.31)$ $(0.15, 0.50)$ $(-0.82, 0.61)$

Table 12. Analytic and simulation results for a regression with varying set of controls on data generated by Case 4.

Model 4 Analysis			
Regression	Analytic Solution	Expected Numeric	Simulated 95% CI ($n = 100$)
$y \sim s$	$\beta_s = l_{sy}$	0.3	$(0.29, 0.31)$
$y \sim x$	$\beta_x = l_{xy}$	0.1	$(-0.16, 0.32)$
$y \sim z$	$\beta_x = l_{zs}l_{sy}$	1.2	$(1.18, 1.23)$
$y \sim x + s$	$\beta_x = l_{xy}$ $\beta_s = l_{sy}$	0.1 0.3	$(0.10, 0.12)$ $(0.30, 0.30)$
$y \sim z + s$	$\beta_x = 0.0$ $\beta_s = l_{sy}$	0.0 0.3	$(-2.3, 1.4)$ $(-0.04, 0.88)$
$y \sim x + s + z$	$\beta_x = l_{xy}$ $\beta_s = l_{sy}$ $\beta_z = 0.0$	0.1 0.3 0.0	$(0.10, 0.12)$ $(0.10, 0.48)$ $(-0.72, 0.79)$

The work presented in this chapter serves as a basis for future work. In our work, we assume full knowledge of the SCM and then proceed with analyzing the consequences of including or excluding different sets of controls in a regression. While it is reasonable to assume that we have full knowledge of the attributes that affect an evaluation score, there is likely to be uncertainty about which attributes have a direct effect on Y. Thus, future work may consider the implications of this uncertainty. Additionally, in our research, we assume that the evaluation process is fixed. In reality, the evaluation process is likely to have changed over time and can be changed in the future. While this may present a challenge, in some respects it could aid in causal inference. Having the flexibility to adjust the evaluation process is equivalent to having the ability to remove or add parent nodes to the evaluation node. Thus, it may be possible to temporarily remove or add edges in a manner that allows for efficient estimation of specific causal quantities of interest. Lastly, the causal models that we consider assume that the attributes are independent. In reality, attributes that are considered during the evaluation process may be correlated and, from a DAG perspective, share a parent node. Analysis of more complex causal models remains an open area of research.

V. Conclusion and Future Work

5.1 Conclusion

In this research, we seek to adapt commercial applications of machine learning and statistics by developing methods that are tailored for Air Force applications through the incorporation of subject matter expertise. In particular, we develop techniques for incorporating subject matter expertise in neural networks, Bayesian regression, and structural causal models. These techniques are developed in the context of three separate application areas:

- Neural networks for localizing point defects in transmission electron microscopy (TEM) of crystalline materials.
- Bayesian regression for estimating the relationship between attributes of fighter pilot communities and flight mishap rate.
- Structural causal models for analyzing Air Force evaluation process.

In Chapter II, we present a novel method, PCA-CNN, for localizing defects in TEM images of crystalline materials. The PCA-CNN method is a self-supervised method that can be trained entirely on TEM images that are free of defects and exhibits strong performance on simulated data. The ability to train a defect localization method without labeled examples of defects represents a novel methodological contribution. Notably, the design of the PCA-CNN method leverages knowledge about point defects in crystalline materials. We show that the tailored design of the PCA-CNN method allows it to outperform CutPaste, a state-of-art, general-purpose defect localization method. Furthermore, we demonstrate the flexibility and generalization performance of the PCA-CNN model by applying it to an experimental image of an unknown crystalline material.

In Chapter III, we first present fighter safety data and trends for class A, B, and C flight mishaps from 2007-2020. Notably, we find evidence of abnormalities in the distribution of mishap cost estimates near the threshold between class B and C mishaps. Thus, we focus on all class A, B, and C mishaps, or HCMs. Next, we quantify pilot attributes and present trends in MDS pilot communities using over 15 years of data. Our analysis reveals trends in the pilot communities of each MDS and suggests there are notable differences between pilot communities. We use these data to model HCM rates as a function of pilot attributes using Bayesian regression and conduct feature selection via predictive projection. By comparing both models with and without personnel factors, we find evidence of a meaningful relationship between personnel factors and HCM rate. Specifically, we find that MDS pilot communities with higher average flight hours in the last year are associated with lower mishap rates. Additionally, we find evidence that higher percentages of pilots who are DGs, are IPs, and have advanced academic degrees are associated with reduced HCM rates. Our efforts to analyze mishap cost estimation behavior, quantify pilot attributes, and model the relationship between pilot attributes and mishap rates represent an applied contribution to the existing literature. Furthermore, our use of Bayesian regression with predictive projection for feature selection represents a valuable methodological contribution. Lastly, we demonstrate the use of Bayesian priors for incorporating the findings of prior qualitative research.

In Chapter IV, we provide a framework for estimating causal effects using data that are associated with Air Force evaluation processes. Unlike in many other applications of data analysis, the causal relationships involved in Air Force evaluation processes are known. Structural causal model provide a mechanism to incorporate this existing knowledge of the evaluation processes. If we assume a linear causal model, we can use the directed acyclic graph defined by the structural causal model

to compute the covariance between any two nodes in the DAG as a function of the causal edge coefficients. We derive a a formula, $\hat{\beta} = C_x^{-1}\vec{C}'_{xy}$, for computing the regression coefficients via the pair-wise covariances of the predictors in the regression. This allows us to express each regression coefficient in terms of the causal edge coefficients in the DAG. Thus, depending on which causal quantity is of interest, we can select the proper controls to ensure that the regression coefficients have a causal interpretation.

5.2 Future Work

There are ample opportunities for future work in each our three contribution areas. In our TEM work, we demonstrated that the PCA-CNN model can be partially trained on experimental data and used to localize simulated defects a single experimental TEM image. With a larger set of experimental data, the PCA-CNN model could be trained exclusively on experimental data. Furthermore, if experimental data with known point defects were available, a more complete evaluation of the generalization ability of the PCA-CNN model woul be possible. Future work could also focus on better understanding the difference in performance of PCA and AI models, such as autoencoders, when analyzing images with repeating structures.

In regards to the fighter safety research, access to higher-fidelity data would be crucial to any future work. Despite our efforts to aggregate multiple data sources and clean existing data, the data quality issues prevented from analyzing fighter safety mishaps at the squadron or group level. If more detailed data were available, quantitative models could be used to estimate the association between factors such as commander rating or unit climate survey results, and fighter mishaps. From a methodological standpoint, the predictive projection method is a highly flexible method for variable selection that warrants further attention. Specifically, the refer-

ence model in the predictive projection method can be nearly any predictive model such as a random forest of neural network. Thus, an existing black-box method could be used as a reference model that is then projected onto a more interpretable submodel.

The use of causal inference theory in analyzing Air Force evaluation processes presents numerous opportunities for future work. First, future work could explore the possibility of iteratively changing an evaluation process to aid in causal effect estimation. Second, more complex causal structures with unobserved variables could be analyzed. Lastly, there are many evaluation processes in the Air Force where the outcome of a selection process is binary. These processes warrant further research because the evaluation process can determine whether future outcomes are observed or unobserved.h.

Bibliography

1. J. Dan, X. Zhao, and S. J. Pennycook, "A machine perspective of atomic defects in scanning transmission electron microscopy," *InfoMat*, 2019.

2. M. Jiang, H. Xiao, S. Peng, L. Qiao, G. Yang, Z. Liu, and X. Zu, "First-Principles Study of Point Defects in GaAs/AlAs Superlattice: the Phase Stability and the Effects on the Band Structure and Carrier Mobility," *Nanoscale Research Letters*, 2018.

3. S. Jesse, M. Chi, A. Belianinov, C. Beekman, S. V. Kalinin, A. Y. Borisevich, and A. R. Lupini, "Big Data Analytics for Scanning Transmission Electron Microscopy Ptychography," *Scientific Reports*, 2016.

4. C.-L. Li, K. Sohn, J. Yoon, and T. Pfister, "CutPaste: Self-Supervised Learning for Anomaly Detection and Localization," 4 2021. [Online]. Available: http://arxiv.org/abs/2104.04015

5. Q. Zhang, M. Zhang, T. Chen, Z. Sun, Y. Ma, and B. Yu, "Recent advances in convolutional neural network acceleration," *Neurocomputing*, vol. 323, pp. 37–51, 2019.

6. F. Wang, T. R. Henninen, D. Keller, and R. Erni, "Noise2Atom: unsupervised denoising for scanning transmission electron microscopy images," *Applied Microscopy*, 2020.

7. S. Mohan, R. Manzorro, J. L. Vincent, B. Tang, D. Y. Sheth, E. P. Simoncelli, D. S. Matteson, P. A. Crozier, and C. Fernandez-Granda, "Deep Denoising For Scientific Discovery: A Case Study In Electron Microscopy," pp. 1–24, 2020. [Online]. Available: http://arxiv.org/abs/2010.12970

8. J. M. Ede and R. Beanland, "Partial Scanning Transmission Electron Microscopy with Deep Learning," *Scientific Reports*, vol. 10, no. 1, 2020.

9. J. A. Aguiar, M. L. Gong, R. R. Unocic, T. Tasdizen, and B. D. Miller, "Decoding crystallography from high-resolution electron imaging and diffraction datasets with deep learning," *Science Advances*, vol. 5, no. 10, 2019.

10. W. Li, K. G. Field, and D. Morgan, "Automated defect analysis in electron microscopic images," *npj Computational Materials*, 2018.

11. M. Ziatdinov, O. Dyck, A. Maksov, X. Li, X. Sang, K. Xiao, R. R. Unocic, R. Vasudevan, S. Jesse, and S. V. Kalinin, "Deep Learning of Atomically Resolved Scanning Transmission Electron Microscopy Images: Chemical Identification and Tracking Local Transformations," *ACS Nano*, vol. 11, no. 12, 2017.

12. J. Madsen, P. Liu, J. Kling, J. B. Wagner, T. W. Hansen, O. Winther, and J. Schiøtz, "A Deep Learning Approach to Identify Local Structures in Atomic-Resolution Transmission Electron Microscopy Images," *Advanced Theory and Simulations*, vol. 1, no. 8, 2018.

13. P. Potapov and A. Lubk, "Optimal principal component analysis of stem xeds spectrum images," *Advanced Structural and Chemical Imaging*, vol. 5, no. 1, 2019. [Online]. Available: https://doi.org/10.1186/s40679-019-0066-0

14. N. Mevenkamp, P. Binev, W. Dahmen, P. M. Voyles, A. B. Yankovich, and B. Berkels, "Poisson noise removal from high-resolution STEM images based on periodic block matching," *Advanced Structural and Chemical Imaging*, 2015.

15. L. Xu, J. Li, Y. Shu, and J. Peng, "SAR image denoising via clustering-based principal component analysis," *IEEE Transactions on Geoscience and Remote Sensing*, 2014.

16. M. Verbanck, J. Josse, and F. Husson, "Regularised PCA to denoise and visualise data," *Statistics and Computing*, 2015.

17. W. Khademi, S. Rao, C. Minnerath, G. Hagen, and J. Ventura, "Self-supervised Poisson-Gaussian denoising," 2020.

18. T. Le, R. Chartrand, and T. J. Asaki, "A variational approach to reconstructing images corrupted by poisson noise," *Journal of Mathematical Imaging and Vision*, 2007.

19. A. Suveer, A. Gupta, G. Kylberg, and I. M. Sintorn, "Super-resolution reconstruction of transmission electron microscopy images using deep learning," in *Proceedings - International Symposium on Biomedical Imaging*, 2019.

20. K. de Haan, Z. S. Ballard, Y. Rivenson, Y. Wu, and A. Ozcan, "Resolution enhancement in scanning electron microscopy using deep learning," 2019.

21. M. Haselmann, D. P. Gruber, and P. Tabatabai, "Anomaly Detection Using Deep Learning Based Image Completion," in *Proceedings - 17th IEEE International Conference on Machine Learning and Applications, ICMLA 2018*, 2019.

22. P. Bergmann, "MVTec AD — A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection," pp. 9592–9600.

23. "National Commission on Military Aviation Safety," Tech. Rep.

24. A. T. Miranda, "Understanding Human Error in Naval Aviation Mishaps," *Human Factors*, 2018.

25. A. R. Gaines, M. B. Morris, and G. Gunzelmann, "Fatigue-related aviation mishaps," *Aerospace Medicine and Human Performance*, 2020.

26. R. T. Nullmeyer, D. Stella, G. a. Montijo, and S. W. Harden, "Human factors in Air Force flight mishaps: Implications for change," *The Interservice/Industry Training, Simulation & Education Conference*, 2005.

27. R. J. Poisson and M. E. Miller, "Spatial disorientation mishap trends in the U.S. Air Force 1993 -2013," *Aviation Space and Environmental Medicine*, 2014.

28. B. J. Hooper and D. P. O'Hare, "Exploring human error in military aviation flight safety events using post-incident classification systems," *Aviation Space and Environmental Medicine*, 2013.

29. J. J. Armentrout, D. A. Holland, K. J. O'Toole, and W. R. Ercoline, "Fatigue and related human factors in the near crash of a large military aircraft," *Aviation Space and Environmental Medicine*, 2006.

30. T. Light, T. Hamilton, and S. Pfeifer, "Trends in U.S. Air Force Aircraft Mishap Rates (1950–2018)," *Trends in U.S. Air Force Aircraft Mishap Rates (1950–2018)*, 2020.

31. D. A. Pamplona and C. J. P. Alves, "Does a fighter pilot live in the danger zone? A risk assessment applied to military aviation," *Transportation Research Interdisciplinary Perspectives*, 2020.

32. T. J. Lyons and W. Nace, "Aircraft crash rates and cumulative hours: USAF data for 25 airframes, 1950-2006," *Aviation Space and Environmental Medicine*, 2007.

33. D. Boyd, "A review of general aviation safety (1984-2017)," *Aerospace Medicine and Human Performance*, vol. 88, pp. 657–664, 2017.

34. J. B. Sobieralski, "The cost of general aviation accidents in the United States," *Transportation Research Part A: Policy and Practice*, 2013.

35. M. Bazargan and V. Guzhva, "Impact of gender, age and experience of pilots on general aviation accidents," *Accident Analysis and Prevention*, vol. 43, pp. 962–970, 2011.

36. D. Rios Insua, C. Alfaro, J. Gomez, P. Hernandez-Coronado, and F. Bernal, "Forecasting and assessing consequences of aviation safety occurrences," *Safety Science*, vol. 111, 2019.

37. C. V. Oster, J. S. Strong, and C. K. Zorn, "Analyzing aviation safety: Problems, challenges, opportunities," *Research in Transportation Economics*, vol. 43, no. 1, 2013.

38. A. Barnett, "Aviation safety: A whole new world?" *Transportation Science*, vol. 54, no. 1, 2020.

39. V. J. Gawron, "Summary of fatigue research for civilian and military pilots," *IIE Transactions on Occupational Ergonomics and Human Factors*, vol. 4, pp. 1–18, 2016. [Online]. Available: https://doi.org/10.1080/21577323.2015.1046093

40. S. Bendak and H. S. Rashid, "Fatigue in aviation: A systematic review of the literature," p. 102928, 3 2020.

41. P. R. Haunschild and B. N. Sullivan, "Learning from complexity: Effects of prior accidents and incidents on airlines' learning," 2002.

42. L. N. Moses and I. Savage, "The effect of firm characteristics on truck accidents," *Accident Analysis and Prevention*, vol. 26, pp. 173–179, 4 1994.

43. F. J. Forteza, J. M. Carretero-Gómez, and A. Sesé, "Occupational risks, accidents on sites and economic performance of construction firms," *Safety Science*, vol. 94, pp. 61–76, 4 2017.

44. J. K. Wachter and P. L. Yorio, "A system of safety management practices and worker engagement for reducing and preventing accidents: An empirical and theoretical investigation," *Accident Analysis and Prevention*, vol. 68, pp. 117–130, 7 2014.

45. P. Baran, P. Zieliński, and Dziuda, "Personality and temperament traits as predictors of conscious risky car driving," *Safety Science*, vol. 142, 2021.

46. T. Nordfjærn, S. Jørgensen, and T. Rundmo, "Cultural and socio-demographic predictors of car accident involvement in Norway, Ghana, Tanzania and Uganda," *Safety Science*, vol. 50, no. 9, 2012.

47. N. Elmitiny, X. Yan, E. Radwan, C. Russo, and D. Nashar, "Classification analysis of driver's stop/go decision and red-light running violation," *Accident Analysis and Prevention*, vol. 42, no. 1, 2010.

48. A. Morgan and F. L. Mannering, "The effects of road-surface conditions, age, and gender on driver-injury severities," *Accident Analysis and Prevention*, vol. 43, no. 5, 2011.

49. G. Li, S. Eben Li, and B. Cheng, "Field operational test of advanced driver assistance systems in typical Chinese road conditions: The influence of driver gender, age and aggression," *International Journal of Automotive Technology*, vol. 16, no. 5, 2015.

50. O. Oviedo-Trespalacios, A. P. Afghari, and M. M. Haque, "A hierarchical Bayesian multivariate ordered model of distracted drivers' decision to initiate risk-compensating behaviour," *Analytic Methods in Accident Research*, vol. 26, 2020.

51. Y. Ali, M. M. Haque, Z. Zheng, and M. C. Bliemer, "Stop or go decisions at the onset of yellow light in a connected environment: A hybrid approach of decision tree and panel mixed logit model," *Analytic Methods in Accident Research*, vol. 31, 2021.

52. Y. Ali, M. M. Haque, and Z. Zheng, "An Extreme Value Theory approach to estimate crash risk during mandatory lane-changing in a connected environment," *Analytic Methods in Accident Research*, vol. 33, 2022.

53. Air Education and Training Command, "T-6a primary pilot training," Joint Base San Antonio–Randolph, Texas: Headquarters Air Education and Training Command, Syllabus P-V4A-J.

54. ——, "T-38c specialized undergraduate pilot training," Joint Base San Antonio–Randolph, Texas: Headquarters Air Education and Training Command, Syllabus P-V4A-J.

55. M. G. Mattock, B. J. Asch, J. Hosek, and M. Boito, *The Relative Cost-Effectiveness of Retaining Versus Accessing Air Force Pilots.* Santa Monica, CA: RAND Corporation, 2019.

56. Department of Defense, "Mishap Notification, Investigation, Reporting, and Record Keeping," *Dodi 6055.07*, no. 6055, pp. 1–52, 2011.

57. J. McCrary, "Manipulation of the running variable in the regression discontinuity design: A density test," *Journal of Econometrics*, vol. 142, no. 2, 2008.

58. L. Abramson, "Sequester has air force clipping its wings," May 2013. [Online]. Available: https://www.npr.org/2013/05/11/183014086/sequester-has-air-force-clipping-its-wings

59. G. Starosta, "The F-35 readies for takeoff," *Air Force Magazine*, vol. 96, no. 4, pp. 38–42, 2013.

60. J. Piironen, M. Paasiniemi, and A. Vehtari, "Projective inference in high-dimensional problems: Prediction and feature selection," *Electronic Journal of Statistics*, vol. 14, no. 1, 2020.

61. T. A. Gormley and D. A. Matsa, "Common errors: How to (and Not to) control for unobserved heterogeneity," *Review of Financial Studies*, vol. 27, no. 2, 2014.

62. J. M. Wooldridge, "Distribution-free estimation of some nonlinear panel data models," *Journal of Econometrics*, vol. 90, no. 1, 1999.

63. P. Guimarães, "The fixed effects negative binomial model revisited," *Economics Letters*, vol. 99, no. 1, 2008.

64. M. L. Blackburn, "The Relative Performance of Poisson and Negative Binomial Regression Estimators," *Oxford Bulletin of Economics and Statistics*, vol. 77, no. 4, 2015.

65. A. Vehtari, A. Gelman, and J. Gabry, "Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC *," Tech. Rep., 2016. [Online]. Available: https://github.com/stan-dev/loo.

66. D. V. Lindley, "The Choice of Variables in Multiple Regression," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 30, no. 1, 1968.

67. J. Piironen and A. Vehtari, "Comparison of Bayesian predictive methods for model selection," *Statistics and Computing*, vol. 27, no. 3, 2017.

68. A. Gelman, J. Hill, and A. Vehtari, *Regression and Other Stories*, 2020.

69. T. Sivula, M. Magnusson, and A. Vehtari, "Uncertainty in Bayesian Leave-One-Out Cross-Validation Based Model Comparison," 8 2020. [Online]. Available: http://arxiv.org/abs/2008.10296

70. "Changes to academic degree and developmental education expectations ¿ air force ¿ article display," https://www.af.mil/News/Article-Display/Article/494376/changes-to-academic-degree-and-developmental-education-expectations/, (Accessed on 04/04/2022).

71. "United States Air Force Science and Technology Strategy: Strengthening USAF Science and Technology for 2030 and Beyond." [Online]. Available: https://apps.dtic.mil/sti/citations/AD1105576

72. T. R. Carretta, "Understanding the Relations Between Selection Factors and Pilot Training Performance: Does the Criterion Make a Difference?" *The International Journal of Aviation Psychology*, vol. 2, no. 2, 1992.

73. D. Schulker, D. Yeung, K. Keller, L. Payne, L. Saum-Manning, K. Curry Hall, and S. Zavislan, *Understanding Demographic Differences in Undergraduate Pilot Training Attrition*, 2018.

74. A. P. Duke and M. J. Ree, "Better candidates fly fewer training hours: Another time testing pays off," *International Journal of Selection and Assessment*, vol. 4, no. 3, 1996.

75. H. Hughes, "Second Language Acquisition: Predicting Less Commonly Taught Languages Learning Success," pp. 41–62, 2018.

76. J. Kling, J. S. Vestergaard, A. B. Dahl, N. Stenger, T. J. Booth, P. Bøggild, R. Larsen, J. B. Wagner, and T. W. Hansen, "Pattern recognition approach to quantify the atomic structure of graphene," *Carbon*, vol. 74, pp. 363–366, 2014.

77. K. M. Keller, S. Robson, K. O'Neill, P. Emslie, L. F. Burgette, L. M. Harrington, and D. Curran, *Promoting Airmen with the Potential to Lead: A Study of the Air Force Master Sergeant Promotion System.* Santa Monica, CA: RAND Corporation, 2014.

78. C. Cinelli, A. Forney, and J. Pearl, "A Crash Course in Good and Bad Controls," *SSRN Electronic Journal*, 2020.

79. E. L. King, D. DiNitto, C. Salas-Wright, and D. Snowden, "Retaining Women Air Force Officers: Work, Family, Career Satisfaction, and Intentions," *Armed Forces and Society*, vol. 46, no. 4, 2020.

80. J. Pearl, "Linear Models: A Useful "Microscope" for Causal Analysis," *Journal of Causal Inference*, vol. 1, no. 1, 2013.

81. D. Poole and M. Crowley, "Cyclic causal models with discrete variables: Markov chain equilibrium semantics and sample ordering," in *IJCAI International Joint Conference on Artificial Intelligence*, 2013.

82. C. Améndola, P. Dettling, M. Drton, F. Onori, and J. Wu, "Structure learning for cyclic linear causal models," in *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence, UAI 2020*, 2020.

83. A. Hyttinen, F. Eberhardt, and P. O. Hoyer, "Learning linear cyclic causal models with latent variables," *Journal of Machine Learning Research*, vol. 13, 2012.

84. S. Cunningham, "Causal Inference: The Mixtape," *Unpublished Manuscript*, 2021.

85. "Is there a way to use the covariance matrix to find coefficients for multiple regression? - cross validated," https://stats.stackexchange.com/questions/107597/is-there-a-way-to-use-the-covariance-matrix-to-find-coefficients-for-multiple-re, (Accessed on 08/01/2022).

86. F. N. David and H. Cramer, "Mathematical Methods of Statistics." *Biometrika*, vol. 34, no. 3/4, 1947.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704–0188

1. REPORT DATE *(DD–MM–YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From — To)*
19-08-2022	PhD Dissertation	Sept 2019 — Aug 2022

4. TITLE AND SUBTITLE

Leveraging Subject Matter Expertise to Optimize Machine Learning Techniques for Air and Space Applications

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Cho, Philip Y., Major, USAF

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

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

Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/EN)
2950 Hobson Way
WPAFB OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT-ENC-DS-22-5-002

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

10. SPONSOR/MONITOR'S ACRONYM(S)

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

12. DISTRIBUTION / AVAILABILITY STATEMENT

DISTRIBUTION STATEMENT A:
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

In this research, we develop machine learning and statistical methods that are tailored for Air Force applications through the incorporation of subject matter expertise. In particular, we develop techniques for incorporating subject matter knowledge in neural networks, Bayesian regression, and structural causal models. These techniques are developed in the context of three separate application areas: localizing point defects in transmission electron microscopy (TEM) of crystalline materials; estimating the relationship between attributes of fighter pilot communities and flight mishap rate; and analyzing Air Force evaluation process.

15. SUBJECT TERMS

Machine Learning, Artificial Intelligence, Bayesian Statistics, Causal Inference

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Dr. Aihua Wood, AFIT/ENC
U	U	U	UU	104	19b. TELEPHONE NUMBER *(include area code)* (937) 255-3636 x4272; aihua.wood@afit.edu