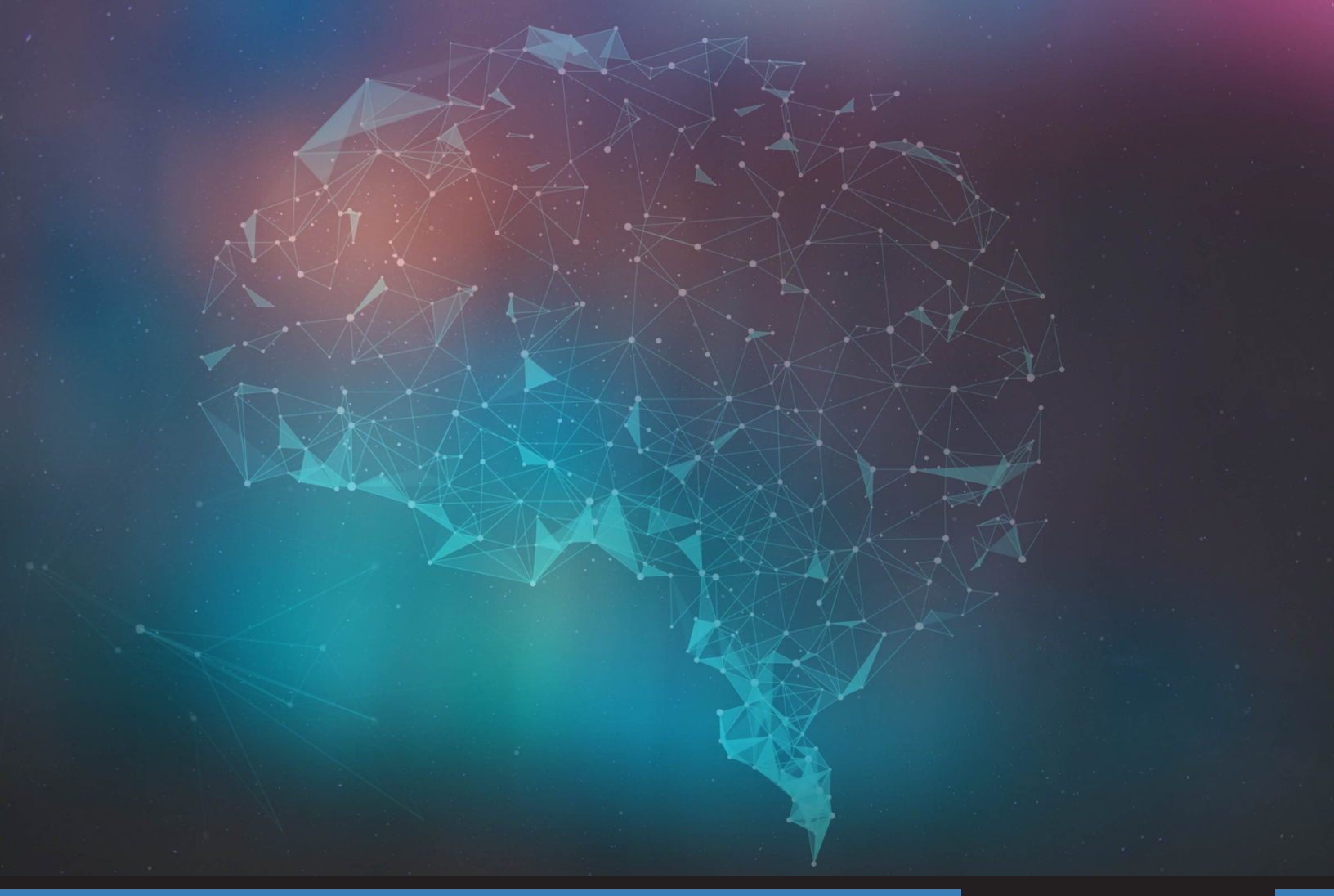


Machine Learning Applications in Nonproliferation:

Assessing Algorithmic Tools for Strengthening Strategic Trade Controls

CNS
Washington, DC Office
NONPRO NOTES
August 2020

Jamie Withorne

Middlebury Institute *of* International Studies at Monterey
James Martin Center for Nonproliferation Studies

James Martin Center for Nonproliferation Studies (Washington, DC Office)
Middlebury Institute for International Studies at Monterey

1400 K Street, NW, Suite 1225, Washington, DC 20005
Phone: +1 (202) 842-3100
www.nonproliferation.org/dc

The author would like to extend a special thanks to Catherine Dill and Jill Luster without whom this research would not have been possible.

Executive Summary

The goal of this report is to demonstrate how machine learning image classification tools may supplement traditional strategic trade controls to bolster nonproliferation efforts.

The author of this report was able to create a successful proof of concept, demonstrating machine learning models' ability to classify controlled dual-use goods from images publicly available online.The research team's goal was to asses the potential utility of machine learning algorithms to classify relevant images as a supplementary tool for detecting potential violations of strategic trade controls. The algorithms discussed in this report are image classification models, or machine learning algorithms that autonomously classify images.

The report details relevant items of interest and selection methodology, a dataset of relevant WMD-related dual-use equipment, and successful image classification model development and evaluation. It also considers the potential wider applicability of machine-assisted identification for nonproliferation efforts including field applications, such as an image dictionary of controlled goods, automating new image classification, and extracting images from additional online sources such as videos.

Because of the criteria that decide which dual-use goods are subject to export controls, it is unlikely that full automation by means of machine learning image classification is currently possible. While image classification models can recognize objects, this research suggests that, often, the models cannot recognize context or item characteristics such as material and size. Thus, while machine learning image classification tools can improve the efficacy of strategic trade control implementation and respective nonproliferation efforts, the research presented here could imply that these tools and applications might not solve all the problems associated with controlling dual-use goods.

Contents

Introduction

The goal of this report is to demonstrate how machine learning image classification tools may supplement traditional strategic trade controls to bolster nonproliferation efforts.

Recent scientific and societal advancements have paved the way for what is colloquially referred to as the machine learning revolution. While the basic foundation for machine learning was established in the middle of the twentieth century, scientists and researchers have refined the algorithms that shape the field of machine learning over the past decade, leading to a drastic increase in the number of potential applications.[1] Machine learning algorithms make up a large percentage of the field of artificial intelligence (AI), and can be defined as "the process by which a computer system, trained on a given set of examples (or data), develops the ability to perform a task flexibly and autonomously."[2]

With this rise in potential applications comes increased global interest in machine-assisted technology and functions across a wide range of subject areas, including international security and defense sectors. Discussions considering the use of machine learning to further weapons of mass destruction (WMD) arms-control and nonproliferation goals have developed in recent years. These discussions not only shape policy formulation and projections, but also begin to accelerate the pace of relevant nonproliferation data analysis to more effectively yield streamlined evidence of proliferation activities.

Nonproliferation involves multiple efforts to combat the spread and/or growth of WMD and related technology that have the potential to threaten US interests. Nonproliferation is complex work done by a variety of organizations and requires an incredibly vast amount of data. However, for purposes of this report, the scope of the research here focuses on nonproliferation efforts by means of strategic trade controls on dual-use goods.

When implementing strategic trade controls, it can be difficult for relevant officials to accurately and efficiently identify individual, controlled items, particularly when items are dual-use in nature. Images of these items are publicly available online, but officials do not necessarily use such resources in screening

1 "A Machine-Learning Revolution," Physics World, March 4, 2019, <https://physicsworld.com/a/a-machine-learning-revolution/>.

2 Ibid.

potential items of concern, nor do they always have the technical expertise necessary to positively identify controlled items.

Machine learning tools can begin to solve some of these problems of strategic trade control applications. Specifically, this report will discuss how image classification models—or machine learning algorithms that autonomously classify relevant images—can be trained to recognize controlled dual-use goods. These image classification tools can assist regulators and improve the efficacy of strategic trade controls for nonproliferation efforts.

This report aims to expand the research on machine learning and potential nonproliferation applications by exploring open-source machine learning image classification tools for WMD-related items (including equipment and/or materials.) It will explore if machine learning image classification tools can be employed alongside existing monitoring mechanisms to improve nonproliferation efforts. The report will detail relevant items of interest and selection methodology, a dataset of relevant WMD-related equipment, and successful image classification model development and evaluation. It will also consider potential wider applicability of machine-assisted identification for nonproliferation efforts including field applications, such as an image dictionary of controlled goods, automating new image classification, and extracting images from additional online sources like videos.

Background

Defining Dual-Use Goods

"Dual-use" is a broad term to refer to any item, technology, or software that has both civilian and military applications. Dual-use goods include items that are components of WMDs as well as items that can be used to manufacture components of WMDs. The multifaceted nature of dual-use goods introduces multiple challenges when attempting to regulate commodity flows. The first challenge is that dual-use goods used in civilian applications often promote development and strengthen economic ties. Understanding the context for how dual-use goods are used is the foundation for managing their risks, and regulators often rely on an

aggregation of information to discern the item's proliferation potential.[3] The second challenge innate to dual-use goods is that the threshold for control of a dual-use good is often based on the technical specifications of the item or its eventual use. When trying to regulate proliferation-sensitive dual-use goods, specialized knowledge is often required to determine the intended application of the good and subsequent legal and regulatory implications. Often those who are implementing controls do not possess this specialized knowledge or background information on items that are particularly sensitive to proliferation.

Figure 1: A Fine Positioning Linear Actuator: An Example of a Dual-Use Item Potentially Difficult to Visually Identify

This 'fine positioning linear actuator' is designed for use in space applications and can be used in missile development. At first glance, however, it can appear to merely be a piece of traditional manufacturing or industrial equipment.[4]

The research in this report seeks to address the challenge in identifying proliferation-sensitive goods. Specifically, it seeks to begin to answer the question: how can machine learning be used to make dual-use good identification easier and more precise for effective regulation? This report will focus on dual-use goods that have been explicitly identified through multilateral export-control regimes as WMD-related equipment or items.

3 Urszula McCormack, Darren Roiser, Robert Edel, Evan Manolios, and Jack Nelson, "Demystifying Dual-Use Goods: From the Chlorine in a pool to the antibiotics we take – what your business should be doing," King & Wood Mallesons, April 18, 2017, <https://www.kwm.com/en/hk/knowledge/insights/demystifying-dual-use-goods-20170418>.

4 "Annex Handbook," Missile Technology Control Regime, 2017, <https://mtcr.info/wordpress/wp-content/uploads/2017/10/MTCR-Handbook-2017-INDEXED-FINAL-Digital.pdf>.

Strategic Trade Controls

International and national frameworks seek to aid nonproliferation efforts of dual-use goods. States use multilateral export-control regimes to coordinate national export controls for dual-use good regulation.[5] These regimes are voluntary and informal bodies that unanimously decide on membership and the lists of items to control, e.g. "trigger lists," which identify dual-use goods used for the purpose of WMD proliferation. There are four multilateral export-control groups that govern most proliferation-sensitive goods: the Nuclear Suppliers Group, the Australia Group, the Wassenaar Arrangement, and the Missile Technology Control Regime.[6]

The data collected for this report drew from the control lists from these four multilateral bodies as well as the United States Commerce Control List (CCL).[7] However, applications of this report's findings seek to inform a wider discussion on strategic trade controls.

Machine Learning Precedents

Demonstrated Advantages of Machine Learning

Machine learning is not a new phenomenon, but the recent development of algorithms that automatically apply complex mathematical calculations to large amounts of data demonstrates the advantage of machine learning in enhancing analysis and work efficiency. Moreover, machine learning is easily accessible to multiple sectors.

The growing amounts of publicly available data, powerful and affordable computational processing, and affordable data storage enable machine learning algorithms to produce fast and accurate results against large, complex datasets. Industries including financial services, health care, government, retail, oil and gas, and transportation apply machine learning processes to improve the efficiency of their work.

---

5 "Multilateral Nonproliferation (Export Control) Regimes and Arrangements," eCustoms, <https://www.ecustoms.com/about-us/visual_trade_compliance_resources/multilateral-nonproliferation-export-control-regimes-arrangements/>.

6 "Multilateral Export Control Regimes," Bureau of Industry and Security of the US Department of Commerce, <https://www.bis.doc.gov/index.php/policy-guidance/multilateral-export-control-regimes>.

7 "Commerce Control List," Bureau of Industry and Security of the US Department of Commerce, <https://www.bis.doc.gov/index.php/regulations/commerce-control-list-ccl>.

The machine learning models discussed in this report rely on deep neural networks. In a deep neural network, neurons—or "bite-sized chunks of information"—are layered together to estimate a given probability.[8] Machine learning models built by deep neural networks algorithms can autonomously learn and make predictions; the researcher only needs to provide a dataset and key parameters. This approach is referred to as "supervised learning" or "supervised classification." Applications of machine learning that use a framework of deep neural networks and supervised learning can take the shape of speech recognition, text recognition, robotics, reasoning, and object recognition. These are only a few examples as there are extraordinarily diverse applications available.

Some definitions are first in order. In this report, a machine learning "model" is a trained algorithm capable of predicting if one object is like another based on the object's features.[9] *Training* a model means teaching the algorithm to make predictions based on the input, whereas *validating* the model means introducing new images with similar features to determine how well the model predicts. This report focuses on a classification model, or a model that predicts discrete values (i.e. is this X item, yes or no?).

This report addresses the issue of dual-use item identification; therefore, it discusses object recognition models, or more specifically, image classification models. Image classification models create deep learning algorithms using computer vision. Computer vision is a computer science term and field of study that enables computers to see and process images in the same way that humans do.[10] Convolution neural networks, or CNNs, are a type of deep learning that have provided the most recent advances in computer vision and image recognition.[11] CNNs rely on a layered neural network, but unlike traditional deep learning neural networks, CNNs do not try to understand the entire image all at once. Rather, the CNN architecture makes its predictions from localized regions in a way that mirrors the human visual cortex. CNN architecture will be described in more detail in the model development section of this report.

---

8 Chris Meserole, "What is machine learning?," Brookings Institution, October 4, 2018, <https://www.brookings.edu/research/what-is-machine-learning/>.

9 "Framing: Key ML Terminology," Google Developers, <https://developers.google.com/machine-learning/crash-course/framing/ml-terminology>.

10 "Computer Vision," Techopedia, February 25, 2019, <https://www.techopedia.com/definition/32309/computer-vision>.

11 "Framing: Key ML Terminology," Google Developers.

Common examples of image classification machine learning tools can be found on Kaggle™.[12] Kaggle is an open-source platform for datasets, data-science training, and machine learning practice. One of their more popular competitions was a challenge to create an image classification machine learning model that could identify if an image contained a dog or a cat.[13] Some other examples of image classification models include using machine learning for image-based cancer detection[14] and facial recognition capabilities.[15]

Machine-Assisted Work in Nonproliferation

In addition to numerous studies on how machine learning could enable transformations in the warfare domain, including nuclear risk and strategic stability,[16] several state entities and international organizations are exploring how machine learning can aid nonproliferation.[17] For example, the United States National Laboratories, including Sandia National Laboratories[18] and Lawrence Livermore National Laboratories[19] are implementing machine learning models to "accelerate the pace of nonproliferation data analysis." Similarly,

12 Kaggle™ is an open-source platform for datasets, data science training, and machine learning practice <https://www.kaggle.com/>.

13 For more on this "challenge," see <https://www.kaggle.com/c/dogs-vs-cats>.

14 Zilong Hu, Jinshan Tang, Ziming Wang, Kai Zhang, Ling Zhang, and Qingling Sun, "Deep learning for image-based cancer detection and diagnosis-A survey," Science Direct, Vol. 83, (November 2018), pp. 134-149.

15 "How does facial recognition work?," Norton Securities, <https://us.norton.com/internetsecurity-iot-how-facial-recognition-software-works.html#:~:text=Facial%20recognition%20is%20a%20way,faces%20to%20find%20a%20match>.

16 Melanie Sisson, Jennifer Spindel, Paul Scharre, and Vadim Kozyulin, "The Militarization of Artificial Intelligence," The Stanley Center for Peace and Security, June 2020, <https://stanleycenter.org/publications/militarization-of-artificial-intelligence/>.

17 For more on warfare applications, see "Artificial intelligence, strategic stability and nuclear risk: Euro-Atlantic perspectives," SIPRI, May 6, 2019, <https://www.sipri.org/news/2019/artificial-intelligence-strategic-stability-and-nuclear-risk-euro-atlantic-perspectives-new-sipri>.

18 Zoe Nellie, Maikael A. Thomas, and Natacha Peter-Stein, "Data Analytics for Nuclear Nonproliferation: Recent Experience at Sandia National Laboratories," US National Nuclear Security Administration, November 1, 2017, <https://www.osti.gov/servlets/purl/1431503>.

19 Jeremy Thomas, "Researchers developing deep learning system to advance nuclear nonproliferation analysis," US National Nuclear Security Administration, August 21, 2018, <https://www.llnl.gov/news/researchers-developing-deep-learning-system-advance-nuclear-nonproliferation-analysis>.

the International Atomic Energy Agency (IAEA) publicly discussed how it is leveraging machine learning for intelligence analysis as well as ways to improve analysts' routine workflows using machine learning.[20]

Machine Learning Model Development and Evaluation

Methodology

The research presented in this section is intended to serve as a successful proof of concept demonstrating the potential benefits of machine learning image classification for dual-use controlled goods. While the findings presented below were not applied to a field case study, potential future field applications will be discussed in the conclusion section of this report.

As previously discussed, the goal of this research is to identify dual-use and controlled goods using a machine-assisted image classification model. In practice, this began with identifying items from multilateral export control regimes' control lists, as well as the US CCL; collecting a dataset of selected images directly related to biological, chemical, and missile technology; building image classification models; and training, tuning, and evaluating these models. The methodology for this process will be further described in the proceeding sections.

Dual-Use Good Dataset and Collection

To begin this research, the research team identified control and trigger lists from the relevant multilateral export control regimes and the US CCL. The research team then sifted through eight control and trigger lists that explicitly discuss dual-use goods. These lists include: Nuclear Suppliers Group Guidelines Part One,[21] Nuclear Suppliers Group Guidelines Part Two,[22] Australia Group Common Control List Handbook Volume

---

20 Brian Ulicny, "Toward a more peaceful world: Using technology to aid nonproliferation," Thomson Reuters Labs, June 13, 2018, <https://blogs.thomsonreuters.com/answerson/toward-a-more-peaceful-world-using-technology-to-aid-nonproliferation/>.

21 "Nuclear Suppliers Group Guidelines Part One," <https://www.iaea.org/sites/default/files/publications/documents/infcircs/1978/infcirc254r14p1.pdf>

22 "Nuclear Suppliers Group Guidelines Part Two," <https://www.iaea.org/sites/default/files/publications/documents/infcircs/1978/infcirc254r11p2.pdf>

One: Chemical Weapons-Related Common Control Lists,[23] Australia Group Common Control List Handbook Volume II: Biological Weapons-Related Common Control Lists,[24] Wassenaar Arrangement List of Dual-Use Goods and Technologies and Munitions List (Volume II),[25] Missile Technology Control Regime Annex Handbook 2017,[26] relevant sections from the International Traffic in Arms Regulations list including the United States Munitions List,[27] and relevant sections from the US CCL.[28]

While reviewing these lists, the research team developed the following selection criteria: 1) If the item is controlled solely based on the type of material used to create it, it is not viable for selection, and 2) If the item does not yield enough high-quality images, it is not viable for selection. Because items, such as chemical development-related items, are often controlled based on their materials composition and technical specifications (such as output volume), in the image-collection process it is impossible to assess with certainty that an image is or is not made of a particular material. While the presented dataset may include pictures of items that are made using uncontrolled materials, such as stainless-steel distillation columns, the research team decided it would be first useful to prove that a machine could learn to recognize selected item shapes. Further research could investigate other types of machine-assisted recognition to detect material of fabrication for control applications. Moreover, machine learning models require large datasets

23 "Australia Group Common Control List Handbook Volume One: Chemical Weapons-Related Common Control Lists," <https://www.dfat.gov.au/publications/minisite/theaustraliagroupnet/site/en/documents/Australia-Group-Common-Control-List-Handbook-Volume-I.pdf>.

24 "Australia Group Common Control List Handbook Volume II: Biological Weapons-Related Common Control Lists," <https://www.dfat.gov.au/publications/minisite/theaustraliagroupnet/site/en/documents/Australia-Group-Common-Control-List-Handbook-Volume-II.pdf>.

25 "Wassenaar Arrangement List of Dual-Use Goods and Technologies and Munitions List (Volume II)," <https://www.wassenaar.org/app/uploads/2019/12/WA-DOC-19-PUB-002-Public-Docs-Vol-II-2019-List-of-DU-Goods-and-Technologies-and-Munitions-List-Dec-19.pdf>.

26 "Missile Technology Control Regime Annex Handbook (2017)," <https://mtcr.info/wordpress/wp-content/uploads/2017/10/MTCR-Handbook-2017-INDEXED-FINAL-Digital.pdf>.

27 "International Traffic in Arms Regulations," <https://www.pmddtc.state.gov/ddtc_public?id=ddtc_public_portal_itar_landing>.

28 "United States Commerce Control List," <https://www.bis.doc.gov/index.php/regulations/commerce-control-list-ccl>.

and many of the items listed on the control lists do not have many images publicly available. While this lack of images may be beneficial for nonproliferation efforts, it presented a challenge in dataset collection.

With this criteria in mind, the research team created a preliminary list of items that might be viable for selection. The research team conducted a cursory search for items on the condensed list, attempting to determine which of the items would most likely yield the highest number of results while considering both quantity and quality of image.

After this search process, the research team identified four items of interest: chemical distillation columns, positive-pressure personnel suits (PPPS), biological safety cabinets, and accelerometers. These items are the class labels that the CNS classification model sought to identify. Distillation columns are a chemical-weapon proliferation risk and controlled by the Australia Group in the Volume I Common Control List Handbook. Both biological safety cabinets and positive-pressure personnel suits can be associated with biological-weapons development and are also controlled by the Australia Group in the Volume II Common Control List Handbook. Accelerometers can be used to develop missile technology for WMD delivery systems, and are controlled by the Wassenaar Arrangement, the Missile Technology Control Regime, and the US CCL.[29]

29 It should also be noted that the accelerometers controlled under the MTCR are "linear accelerometers designed for use in inertial navigation systems or in guidance systems of all types….". The research team focused on the broader category of mems accelerometers. Mems accelerometers are the newest type of accelerometer, and while accelerometers can make various different types of measurements, the square mems accelerometers measures x, y, z axis acceleration as opposed to vibration or voltage. While this is a relatively broad category, not accounting for specific brands or versions, and has the potential to pose a difficulty for a classification algorithm due to visual similarities to circuit boards, accelerometers proved to be one of the least problematic items for dataset creation.

Figure 2: Images of the Selected Items[30]

(L–R): Accelerometer, Biosafety Cabinet, Distillation Column, and Positive-Pressure Personnel Suits

The research team also sought to include a nuclear-related item, either from the Nuclear Supplier Group trigger list or the US CCL. While the research team briefly considered selecting baffle plates as an item, there were not enough available images to make this a viable selection option. This difficulty in selecting specific items held true for many of the items listed in control lists. A number of the available images of baffle plates were diagrams, or images with artificial backgrounds, and the research team determined this type of image would not be good training data for a model since it would likely result in overfitting. Overfitting is when a model is so tuned to the training data that it is unable to recognize or classify new images. Overfitting will be further discussed in the model evaluation section of this report.

After item selection, the research team collected images by searching various publicly available sources. Sources included: official governmental and institutional reports, reports on malfunctioning equipment that included selected items, manufacturing reports, distribution sites and inventories, and various analyses from non-governmental organizations and individuals. The research team also conducted an in-depth search of museums and other public displays of the selected items, as well as information from the producers of the items themselves, including various website galleries and image stills from online promotional videos. The research team collected as many images of the four selected items as possible, emphasizing a diverse range of images, with regards to both image size and file type, and aimed to collect

30 Accelerometer image source: <https://www.amazon.in/embsys-MEMS-Accelerometer-Sensor-ADXL335/dp/B00XR0OW5E>; Biosafety Cabinet image source: <https://www.indiamart.com/proddetail/biosafety-cabinet-class-iii-16467740830.html>; Distillation Column source: <https://www.indiamart.com/proddetail/distillation-column-20483547048.html>; Positive-pressure personnel suit image source: <https://commons.wikimedia.org/wiki/File:Positive-pressure_suit_(orange_suit).jpg>.

approximately 200+ images of each item. The research team emphasized images of the selected items isolated or alone, selected items in an operational setting, and selected items with other various objects in the image. The diversification of image type, size, and quality, as well as a large dataset had the aim of improving image classification model results.

Due to the items' dual-use and controlled nature, as well as an emphasis on collecting a diverse image dataset, the collection process was not without limitations, and ultimately no more than 300 images were collected for each item. While these limitations posed challenges for dataset creation, they arguably improve nonproliferation efforts by limiting public access to controlled item parts that could be used for WMD development.

After the selected item image collection, the research team collected negative images to train the machine learning models. Negative images are images that do not contain pictures of the selected items.

The research team created an image dataset that includes: 200 images of accelerometers, 202 images of distillation columns, 97 images of biological safety cabinets, 280 images of positive-pressure personnel suits, and 126 images of items that contained both biological safety cabinets and positive-pressure personnel suits. The research team also collected 201 negative images to train the models on for a total dataset size of 1,106 images. The number of items per class in a dataset is crucial to understanding the performance of the final models, along with associated model metrics and evaluation. Therefore, it is important to understand the dataset before discussing model architecture and evaluation.

Because of the small size of this dataset, the research team briefly considered applying a one-shot learning categorization approach.[31] However, due to the research team's assessed skillsets, this approach was deemed to be out of the scope of this report. Future research could apply a one-shot learning categorization method.

Following image collection, the research team organized each class into "train" and "test" folders. The research team placed 80% of the total number of items in a class into a training folder, and 20% of the total number of items in a class into the test folder. Because of the different number of images in each class, all the training and test folders were composed of a different number of images. The presented datasets are

---

31 Harshall Lamba, "One Shot Learning with Siamese Networks using Keras," Medium, January 21, 2019, <https://towardsdatascience.com/one-shot-learning-with-siamese-networks-using-keras-17f34e75bb3d>.

"unbalanced," and can tend to show a preference toward the class with more representation, or more images. However, these unbalanced datasets do not necessarily affect the accuracy of the model, as will be described in the model evaluation section.

After creating class training and testing datasets, the research team chose to utilize a "flow from dataframe" data upload option. This required building a dataframe readable by the Python™ tool Pandas™ for model development. Python is a general-purpose programming language that the research team relied on for the coding aspects of this research.[32] To create a dataframe, the research team placed the image dataset in their local machine's Python working directory. Separately, the research team created a CSV file that includes the following columns: dataset, filename, directory name (to show the location of the file), "BSC" or biosafety cabinet class, "PPPS" or positive-pressure personnel suit class, "ACC" or accelerometer class, "DC" or distillation column class, and "Negative" for a negative image class. This is a custom class structure that enables the model to classify items into categories relevant to this research's outlined goal. Using the os.walk function in Python, the research team populated the filename and directory name columns to ensure accuracy. The remainder of the dataframe was manually entered, including manually entry if the image fell within a test or train dataset and doing "one hot" encoding for the various class columns. One hot encoding includes entering a one in the column of the class if the image contains said class and a zero in the column of the class if the image does not contain that class. Because the dataset includes images that include both biosafety cabinets and positive-pressure personnel suits, the CNS image recognition model must be a multi-class recognition model, as will be further discussed in the model development section.

Model Architecture, Output, and Evaluation

Model Architecture

At the outset of this research, the research team had limited knowledge of coding, Python™, and creating machine learning models. To begin the model-building process, the research team first familiarized themselves with the necessary tools associated with model building and development. The research team completed several Python language tutorials, the "Machine Learning Crash Course"[33] and "ML Practicum:

32 Python is a general-purpose programming language <https://www.python.org/>.

33 "Machine Learning Crash Course," Google Developers, <https://developers.google.com/machine-learning/crash-course>.

Image Classification"[34] courses by Google Developers™. Additionally, the research team utilized open-source code platforms such as "Stack Overflow"[35] and "GitHub"[36] to inform model development.

Following an initial learning period, the research team installed Python and respective foundational coding packages such as Os™, Tensorflow™, Keras™, Pandas™, Numpy™, SKLearn™, and Matplotlib™, on a CNS local device. The research team relied on the Spyder™ Python environment, a tool through the Anaconda™ open-source Python platform, for model development. Keras™ and TensorFlow™ are the most popular applied programming interfaces that use Python packages for machine learning image classification models.

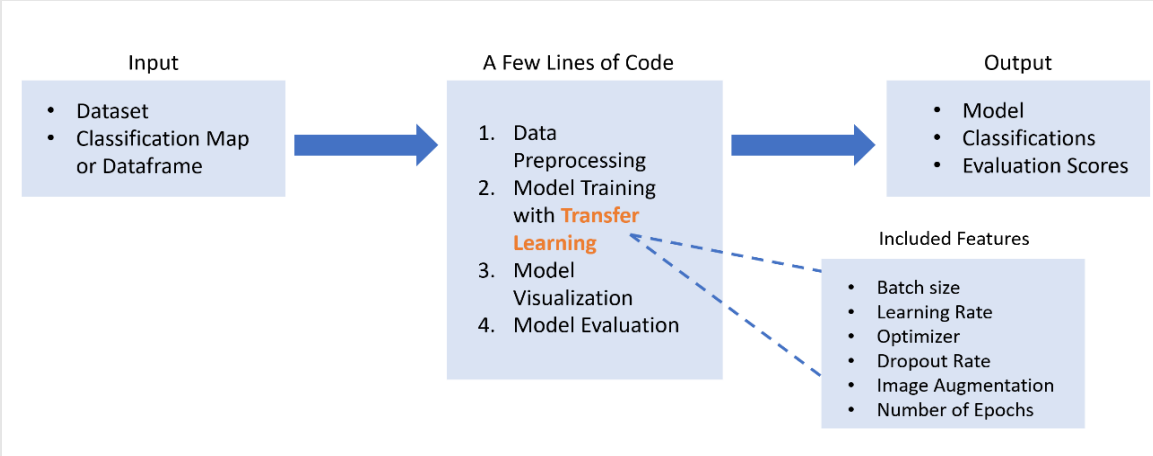
Creating a machine -learning model includes inputting a dataset, coding model specifications, and receiving an output. Model-building methodology includes: uploading a dataset and custom classification system, using code to translate the image data into a format that a machine can understand, and defining the output layers, or characteristics of the model.

34 "Machine Learning Practice: Image Classification," Google Developers, <https://developers.google.com/machine-learning/practica/image-classification/next-steps>.

35 Stack Overflow is an online community for developers to learn and share their programming knowledge <https://stackoverflow.com/>.

36 GitHub is an online community of developers with the aim of discovering sharing, and building better software <https://github.com/>.

Figure 3: Model Development Methodology[37]



In the initial stages of model development, the research team decided that it would be best to create a model that relied on a "flow from dataframe" approach to upload the images. In order to utilize Keras™ on a local device with a customized classification system and dataset, there are two popular formats for dataset upload: flow_from_directory and flow_from_dataframe. Flow_from_directory is more common and organizes the images within separate folders named after their respective class and dataset.[38] Conversely, flow_from_dataframe, involves mapping all the image classes in a CSV or JSON file. Flow_from_dataframe requires the Pandas™ Python™ package to read the class map CSV file and translate it into something a machine can read.

After experimentation with both approaches, the research team found the flow_from_dataframe option to be a more suitable option, since flow_from_directory did not handle multi-label images well. In order to implement this option, after creating the dataframe as previously described, the research team reorganized the folders to have all classes in either a test or train main folder. Similarly, the research team decided to employ a generator function in the foundational code and model architecture because machine learning relies on multiple iterations to learn. Generators behave like an iterator, or code that does iterations on

37 Model building flowchart adopted from Margaret Maynard-Reid, "An Icon Classifier with TensorFlow Lite Model Maker," Medium, May 9, 2020, <https://medium.com/swlh/icon-classifier-with-tflite-model-maker-9263c0021f72>.

38 J. Vijayabhaskar, "Tutorial on Keras flow_from_dataframe," Medium, September 21, 2018, <https://medium.com/@vijayabhaskar96/tutorial-on-keras-flow-from-dataframe-1fd4493d237c>.

data, looping through elements of an object, such as specific areas of an image, and can be used similarly to an array.[39] Generators use less space on devices because they do not hold results in memory, and because of this, it may take longer to run a model without a generator. Because of computational space limitations a generator approach was the most viable option for the CNS developed models.

In the foundational code, pre-processing refers to "the transformation applied to data before feeding it to the (machine learning) algorithm."[40] Simply, pre-processing is a technique that is used to convert raw data, or the images that compose the dataset, into a clean dataset that an algorithm can understand. The function ImageDataGenerator was used in model development in order to pre-process the images for Keras™. The pre-processing included data augmentation which "artificially boosts the diversity and number of training examples by performing random transformations to existing images to create a set of new variants."[41] Data augmentation is a common practice and serves as a preventative measure against overfitting the model. Data augmentation increases data volume and data diversity and is a useful technique to implement.[42] Image augmentation was not used on the validation or test dataset. While testing models using augmented images can be done with a technique referred to as test time augmentation, this approach is often used for model selection, as opposed to image classification, and was not applied in this research.[43]

The research team developed four different image classification models in order to compare results between different approaches in model training and validation. All four models relied on the aforementioned data pre-processing and had a batch size of thirty two for the generator iterations. All four

---

39 Jessica Yung, "Using generators in Python to train machine learning models," jessicayung.com, October, 2018, <https://www.jessicayung.com/using-generators-in-python-to-train-machine-learning-models/>.

40 "Data Preprocessing for Machine learning in Python," Geeksforgeeks.org, <https://www.geeksforgeeks.org/data-preprocessing-machine-learning-python/#:~:text=Data%20Preprocessing%20for%20Machine%20learning%20in%20Python,into%20a%20clean%20data%20set>.

41 "ML Practicum: Image Classification, Preventing Overfitting," Google Developers, <https://developers.google.com/machine-learning/practica/image-classification/preventing-overfitting>.

42 Image augmentation approaches used in this report included image rotation, width and height shifts, shearing, zooming, and horizontal flipping.

43 For more on how to use a Test-Time Augmentation, see Jason Brownlee, "How to Use Test-Time Augmentation to Make Better Predictions," Machine Learning Mastery, April 3, 2020, <https://machinelearningmastery.com/how-to-use-test-time-augmentation-to-improve-model-performance-for-image-classification/>.

models also relied on an Adam optimizer with a learning rate of .0001. An optimizer is required to compile Keras models and determines the rate at which the model learns. The research team initially explored models with different optimizers, learning rates, and batch sizes, but it found this combination of model features to yield the most accurate results. It should be noted that batch sizes smaller than sixteen tend to more readily overfit the model, and a learning rate of .0001 is standard in machine learning models utilizing transfer learning. The research team attempted to do a step decay learning rate on some models during model development in order to see how this would affect model accuracy.[44] Results of the step decay learning rate were negligible. The research team tested the base model using all the available Keras™ optimizers,[45] and found the Adam optimizer to consistently yield the best results.

For model architecture, the research team utilized dropout regularization in the model code to prevent overfitting. Dropout regularization "randomly removes units from the neural network during a training gradient step."[46] Additionally, because the dataset included multi-class images, the research team relied on a sigmoid rather than a softmax output that utilized a binary crossentroypy loss function.[47]

As previously noted, in order to train a machine learning algorithm, a test, train, and validate dataset are needed. The train dataset is used to initially teach or train the model. The validation dataset is used to introduce new images to see if the model can easily recognize new images, or if it is overfit and not able to classify new images. The test dataset is used to evaluate the overall model performance.

Figure 4: Example of Model Training, Validation, and Testing Dataset Processes and Interactions[48]

---

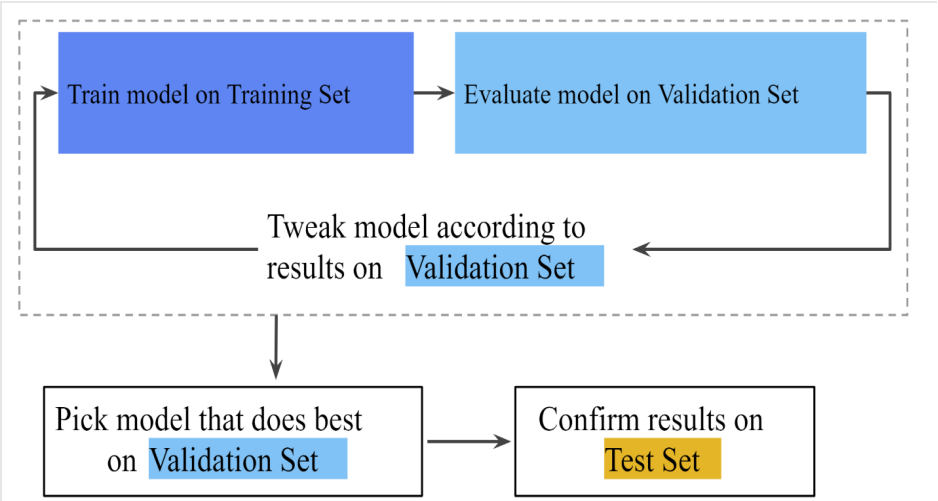
44 A step decay learning rate is a learning rate scheduler that reduces the learning rate during model training according to a pre-defined schedule.

45 Zhijian Li, "Comparison of Optimizers for Keras," Kaggle, 2018, <https://www.kaggle.com/c/human-protein-atlas-image-classification/discussion/70253>.

46 Ibid.

47 A sigmoid or softmax function is put at the end of a neural network classifier to convert raw output values into probabilities. Because there are images with two classes, the research team used a sigmoid output for the models as it is independent and not constrained to sum to one, as opposed to a softmax output. Because the sigmoid layer only allows for a single input value, it requires a binary cross entropy loss function. A sigmoid output layer and binary crossentropy allow for labeling multi-class images.

48 "ML Practicum: Image Classification, Another Split," Google Developers, <https://developers.google.com/machine-learning/practica/image-classification/preventing-overfitting>.



The variables across the four models the research team created include differing validation approaches and using a pre-trained versus a simple CNN model architecture. As the previous discussion on the dataset has demonstrated, the research team initially did not include a validation folder in the data organization. For small datasets such as the one described in this report, it is often recommended to do a k-fold cross validation.[49] After extensive research regarding feasibility, it was determined that implementing a k-fold cross validation alongside generators was out of the scope of the research team's demonstrated ability. This is a validation approach that could be implemented in future research. Because a k-fold cross validation was not feasible, the research team decided to test two different validation approaches: one that automatically took the last 20% of the training images per class and created a validation folder using the validation_split function in Python™, and one that involved manually, randomly selecting 20% of training images from each class and sorting them into a validation folder.

The research team also sought to compare the differences between pre-trained models and models that use a basic, self-created, CNN architecture. A pre-trained model is a model previously trained on larger datasets of a similar problem and, through a process called transfer learning, can improve model accuracy.[50]

---

49 Cross-validation is a resampling procedure that is used to estimate the ability of a model created with a limited dataset to classify unseen data. The process has a parameter, k, that refers to the number of groups a dataset will be split into, hence the name "k-fold."

50 Pedro Marcelino, "Transfer learning from pre-trained models," Medium, October 23, 2018, <https://towardsdatascience.com/transfer-learning-from-pre-trained-models-f2393f124751>.

Transfer learning is the process when a model is trained on a separate task, and those weights are then used as a starting point for the new task.[51] This is a common technique in training deep convolutional neural networks, especially with high performing classifiers. The model and the weights that are used are called a pre-trained model. For the pre-trained model, the research team relied on the InceptionV3 model previously trained on ImageNet from Keras™ that has acquired accuracies over 78%.[52] ImageNet is a large image database, containing several million images, designed for image recognition research.[53]

As described earlier, CNNs receive input feature maps from the image pre-processing stage and create a stack of modules, each of which then performs three operations: convolution (or feature extraction), rectified linear unit transformation to the convolved feature, and pooling where the CNN down-samples the convolved features and reduces the number of dimensions on the feature map.[54] Figure 5 below illustrates the structure of a convolutional neural network.[55]

51 Weights are the learnable parameters of a machine learning model and help to determine the strength of connection between two neurons.

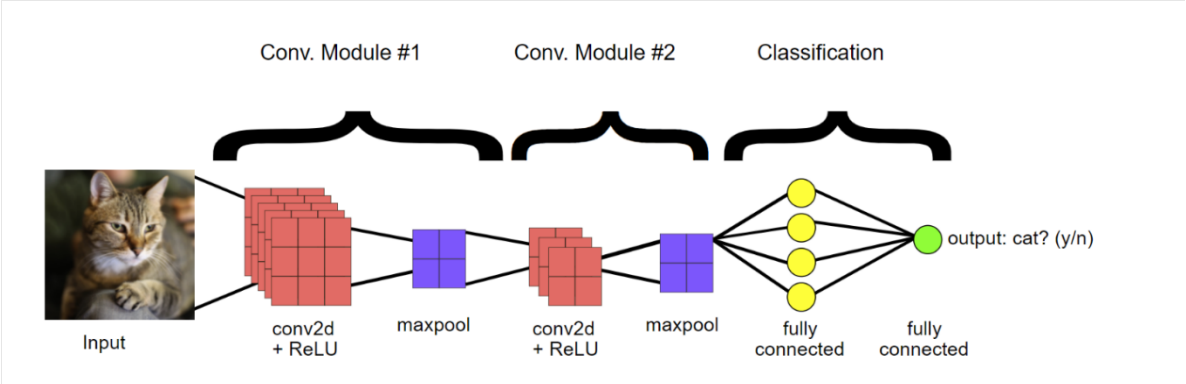
52 Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wonjna, "Rethinking the Inception Architecture for Computer Vision," Cornell University, December 11, 2015, <https://arxiv.org/abs/1512.00567>.

53 ImageNet is a large online image database <http://www.image-net.org/>.

54 "ML Practicum: Image Classification, Introducing Convolutional Neural Networks," Google Developers, <https://developers.google.com/machine-learning/practica/image-classification/convolutional-neural-networks>.

55 For this report's specific CNN architecture, contact the author directly.

Figure 5: A typical CNN structure[56]



The four CNS models varied as follows: Model One used a manual validation and a simple CNN architecture; Model Two used a manual validation and a pre-trained model architecture; Model Three used an automatic validation split and a CNN architecture; and Model Four used an automatic validation split and a pre-trained model architecture.

Figure 6: CNS Models and Respective Training Approaches

Model One Manual Validation and CNN Architecture	**Model Two** Manual Validation and Pre-Trained Model
Model Three Automatic Validation and CNN Architecture	**Model Four** Automatic Validation and Pre-Trained Model

Each model was trained with fifty epochs.[57] All models were set to measure the accuracy with which the machine learning model could classify images. Accuracy scores in classification model evaluations divide

56 Ibid.

57 An epoch is one complete presentation of the dataset to be learned to a learning machine.

the total number of predictions the model made by the number of correct predictions the model made. It should be noted that small jumps or spikes in accuracy are not immediately concerning, as they represent images with which the model may be having more or less difficulty. Accuracy alone does not provide the full picture of how well a model is doing, especially with class-imbalanced datasets, such as the datasets described in this report. Future research could evaluate models using precision and recall metrics for a more complete analysis.

The loss metrics presented below solely represent how far, on average, the model's prediction was from the actual example. As previously mentioned, the models outlined in this report are implementing a binary crossentropy loss function that is not weighted.[58] Loss metrics do not provide much comparative value between models, but rather, are presented to demonstrate learning, or prediction, paces. Large, individual spikes are not immediately concerning since there is an element of randomness in training a model. It is typical that the model performs better on training data than the validation data, but an upward curve in accuracy and a generic downward trend in loss is ideal.

Model Results and Evaluation

Using this described model-building approach, the research team was able to successfully create and train models that were able to recognize and classify dual-use images from the CNS dataset. Figures 7-18 below provide a comparison of training and validation results for the metrics of accuracy and loss for all four CNS models. Training datasets are represented by a blue line, and validation datasets are represented by an orange line. Again, these models were successful for the purposes of this research, and while the metrics may not be perfect, they are nonetheless indicative of a machine that learned to classify dual-use goods and did so effectively.[59]

---

58 While there are other types of loss functions, such as categorical crossentropy, binary crossentropy is best suited for the models described in this report as the models are designed to recognize multi-class images. The underlying mathematical equation of binary crossentropy is: BCE(t,p) = -(t*log(p) + (1-t) *log(1-p)). For more on binary crossentropy, see "How to use binary & categorical crossentropy with Keras?," Machine Curve, October 22, 2019, <https://www.machinecurve.com/index.php/2019/10/22/how-to-use-binary-categorical-crossentropy-with-keras/#binary-crossentropy-for-binary-classification>.

59 For this report's numeric results, contact the author directly.

Figure 7: Model One Accuracy

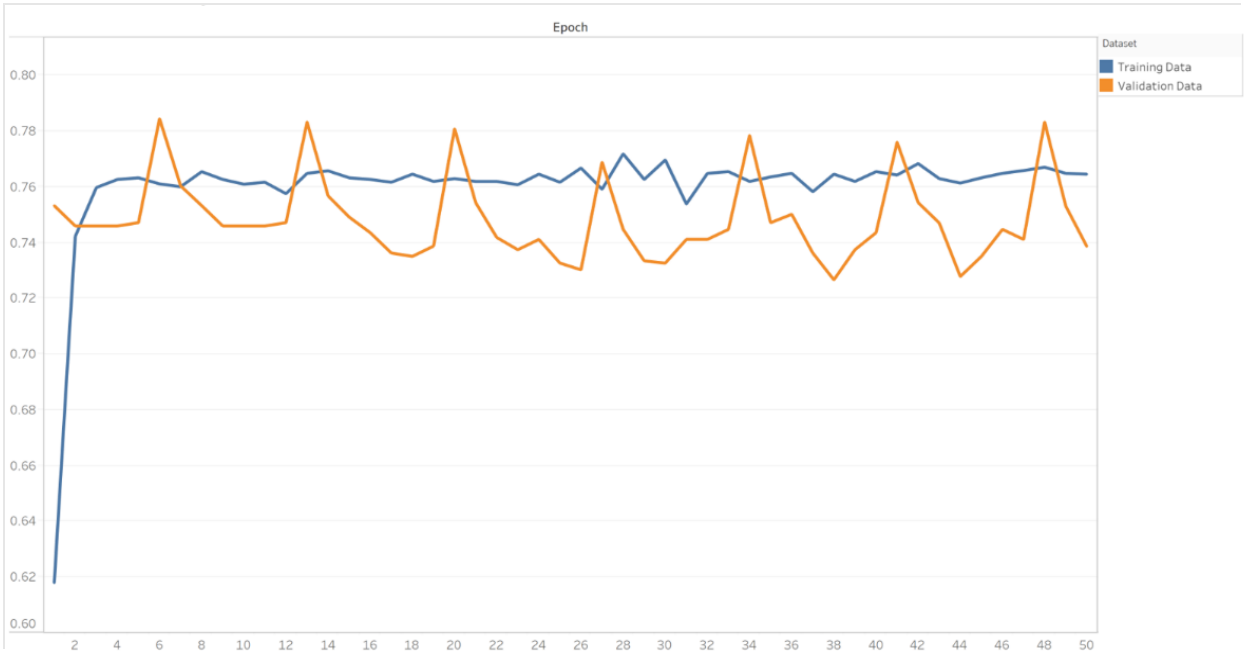


Figure 8: Model One Loss

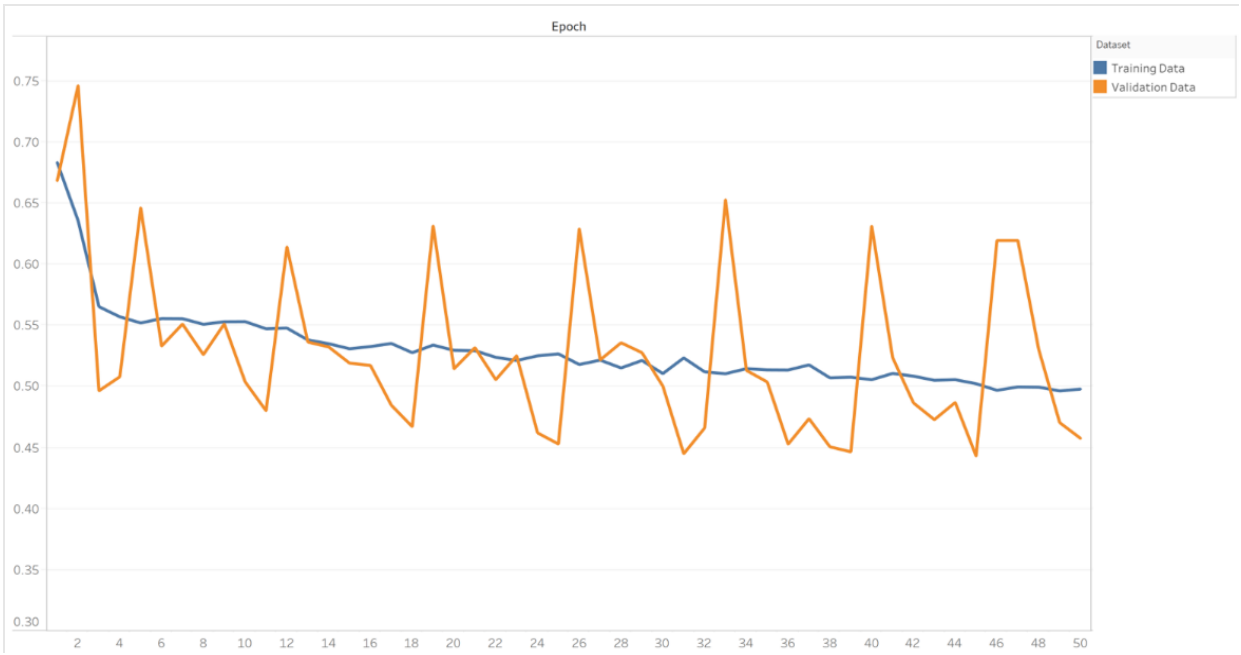


Figure 9: Model Two Accuracy

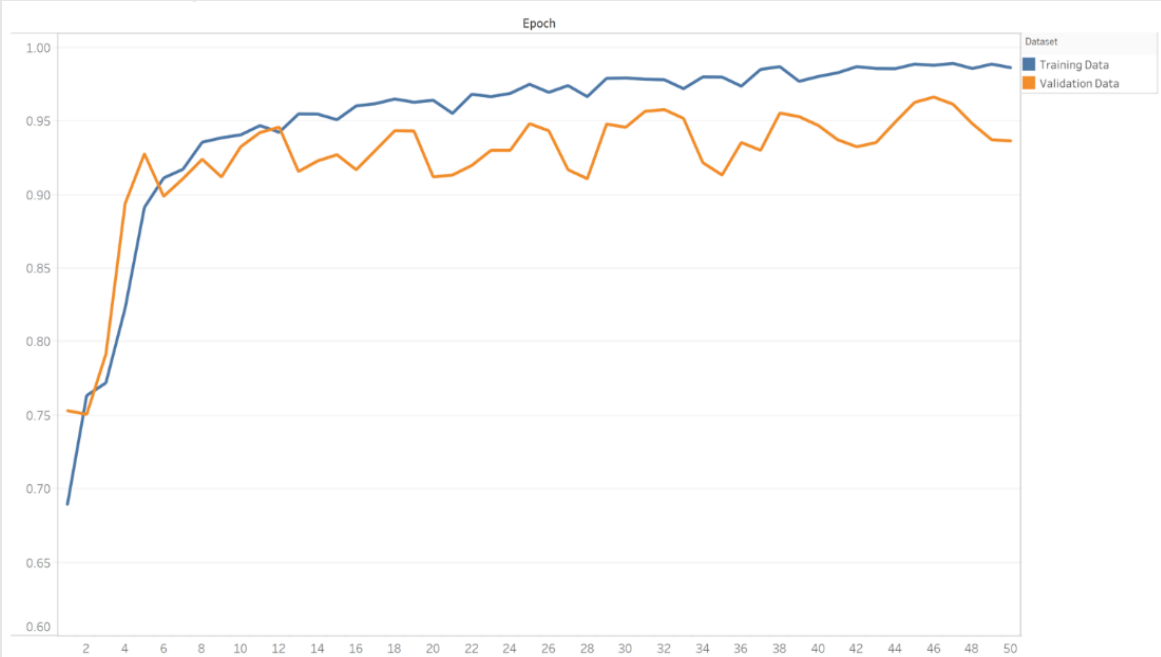


Figure 10: Model Two Loss

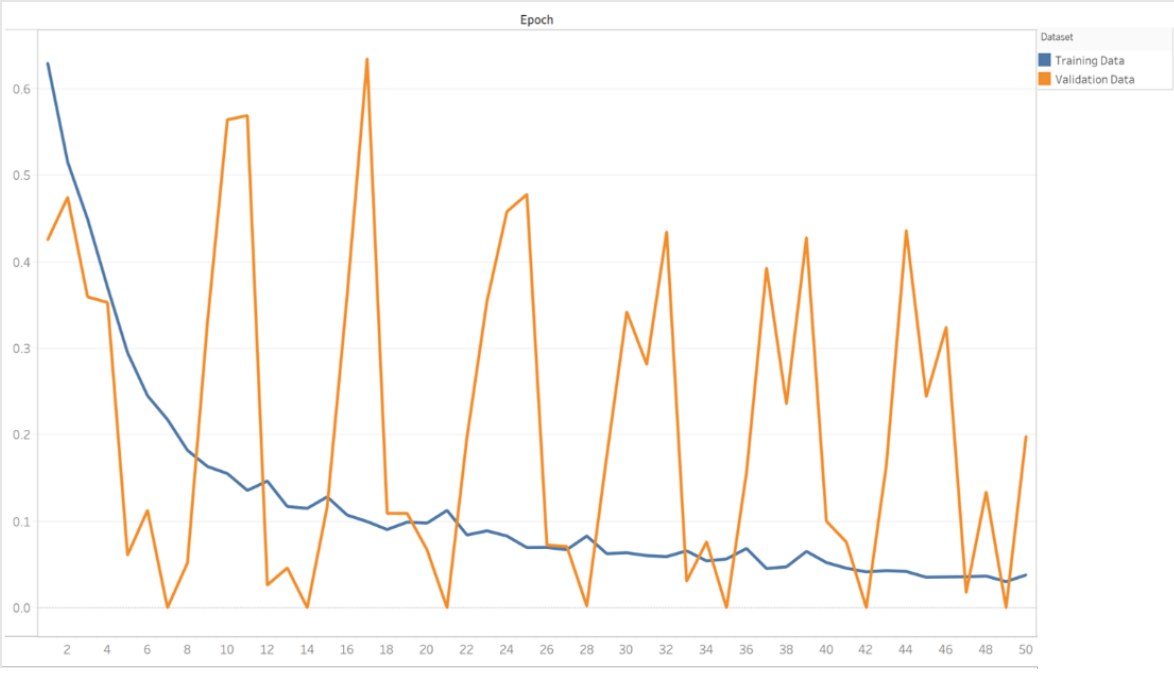


Figure 11: Model Three Accuracy

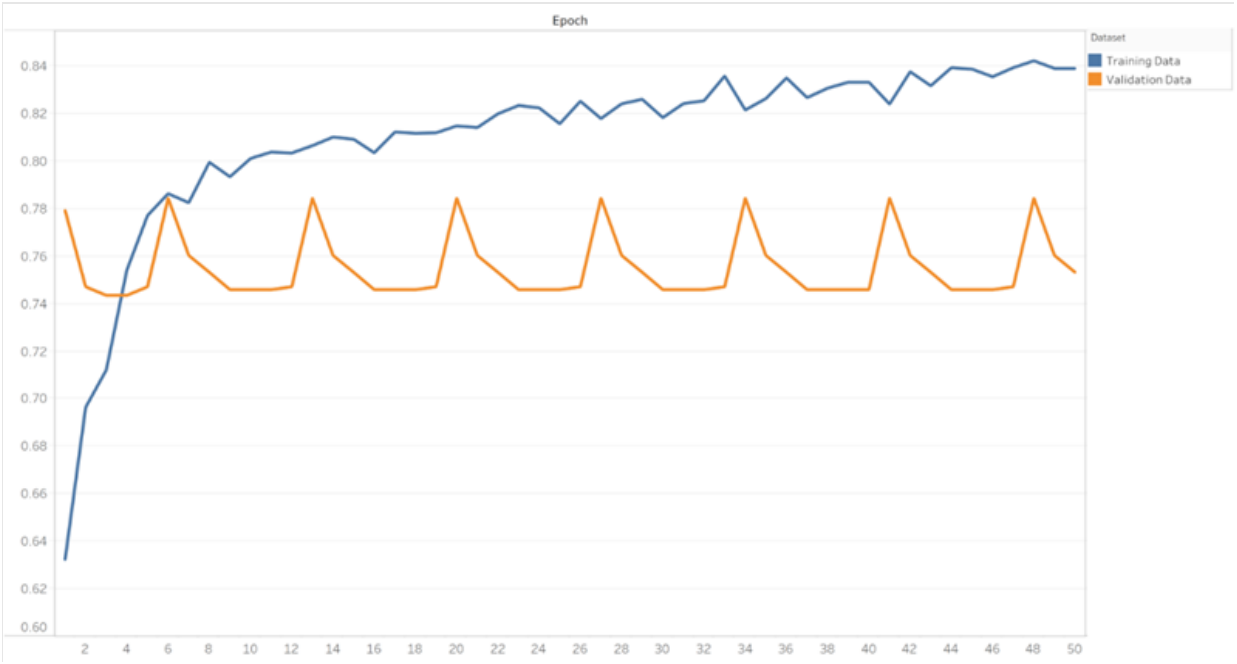


Figure 12: Model Three Loss

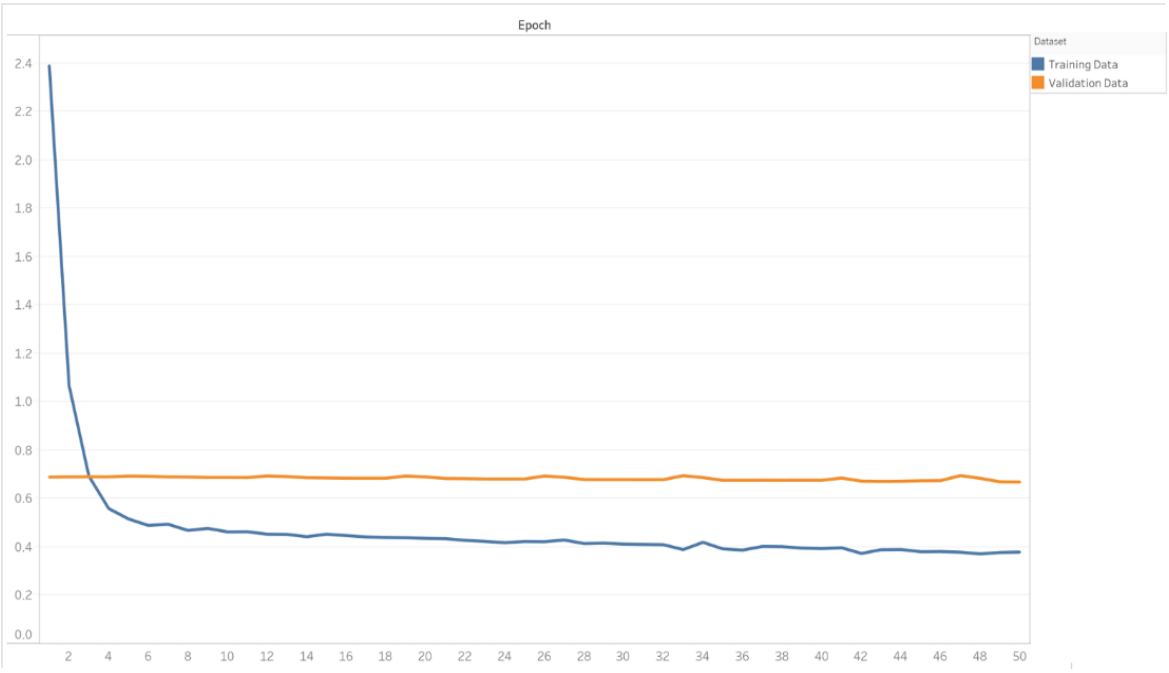


Figure 13: Model Four Accuracy

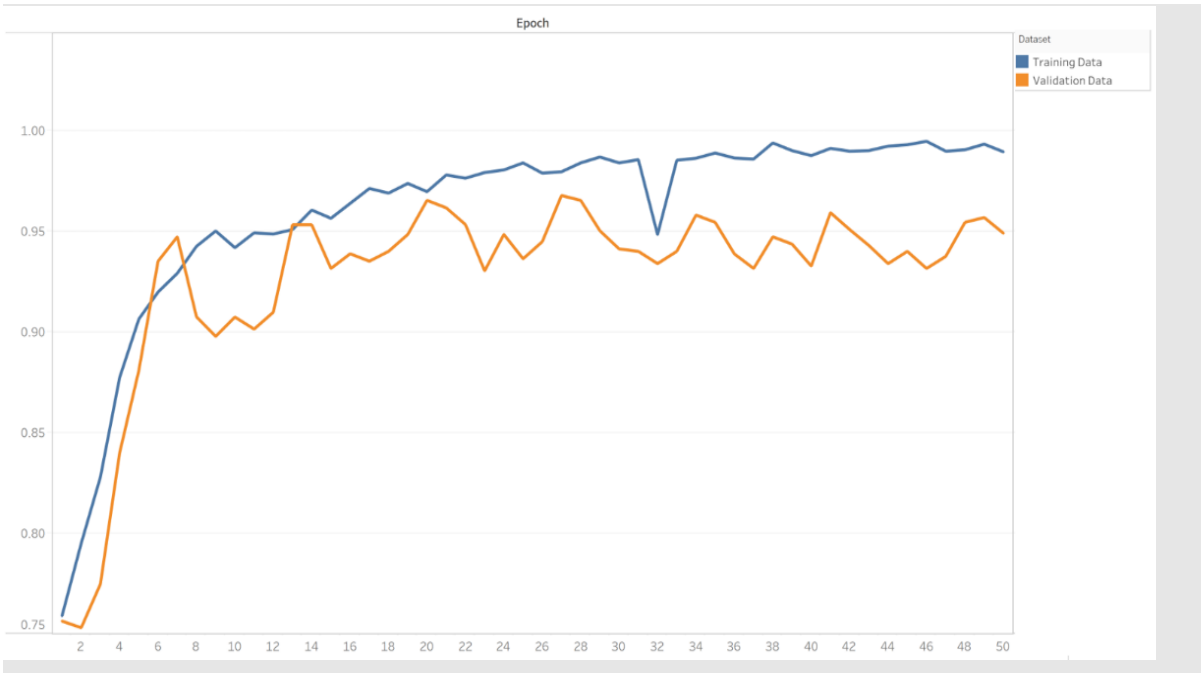


Figure 14: Model Four Loss

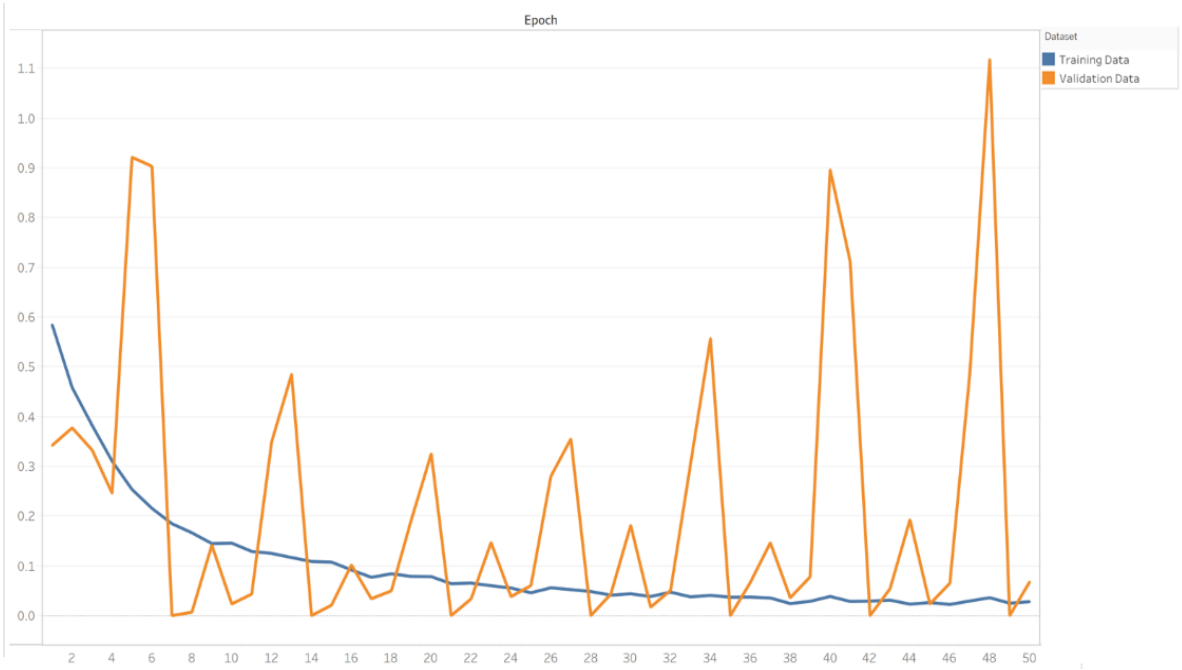


Figure 15: Model Comparison: Training Accuracies

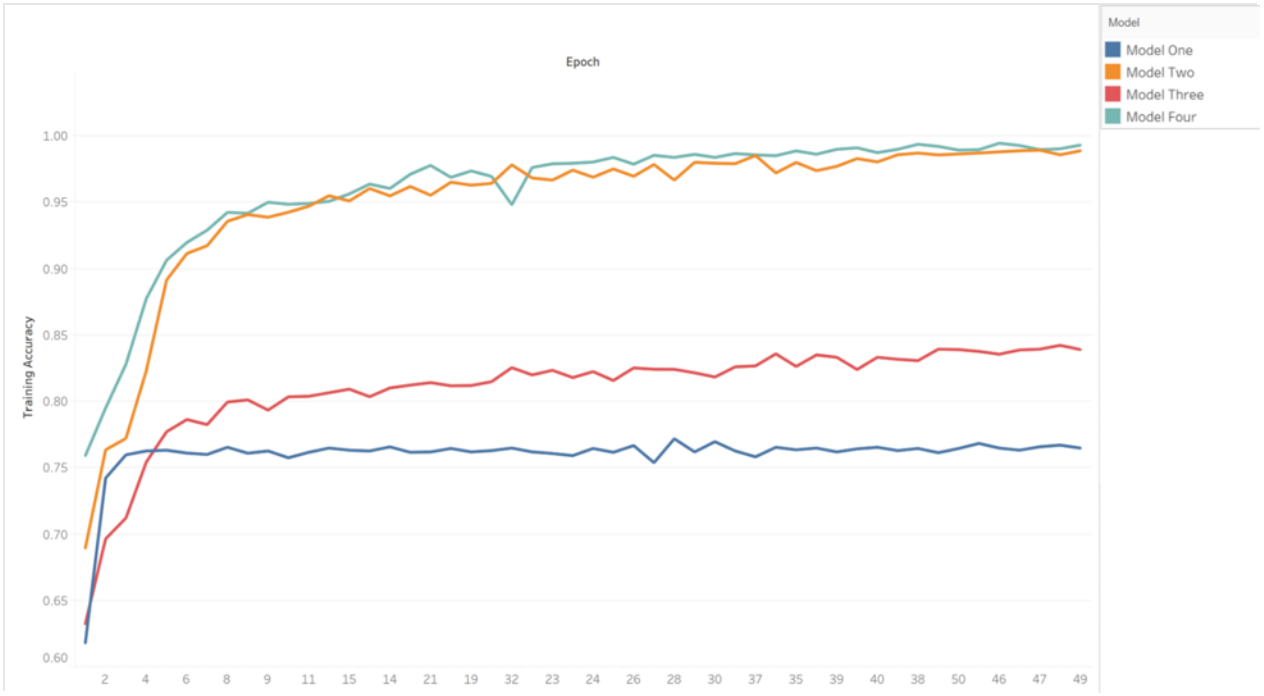


Figure 16: Model Comparison: Validation Accuracies

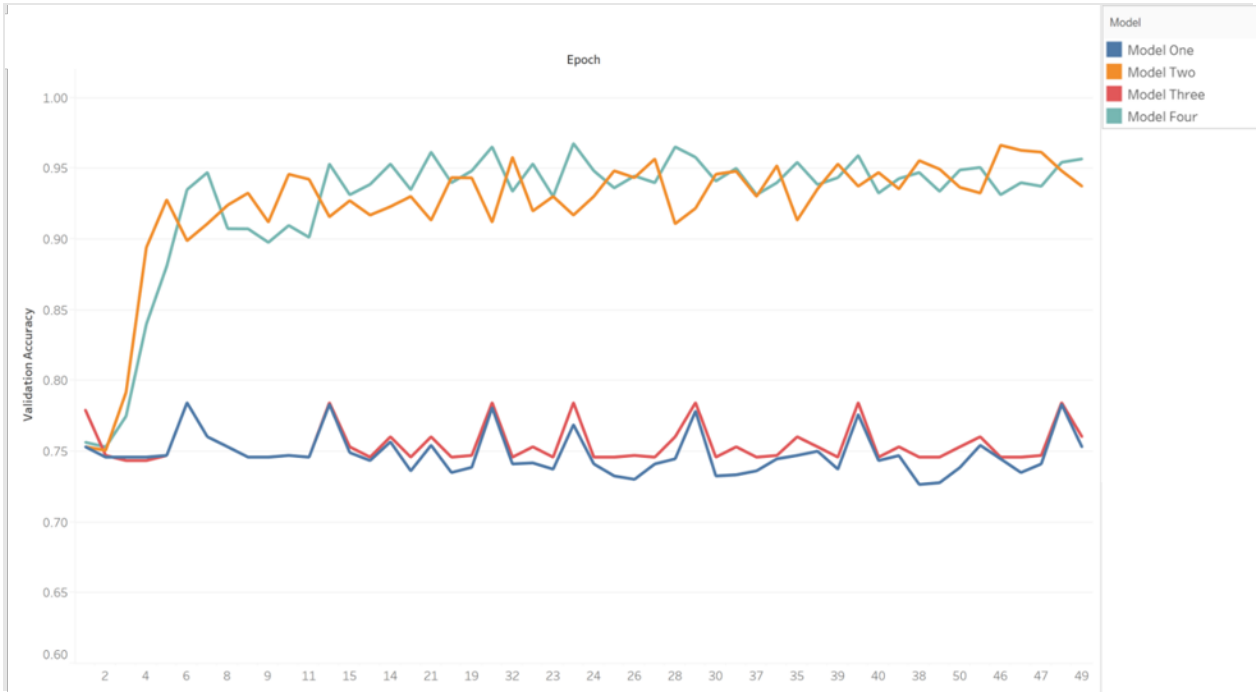


Figure 17: Model Comparison: Training Losses

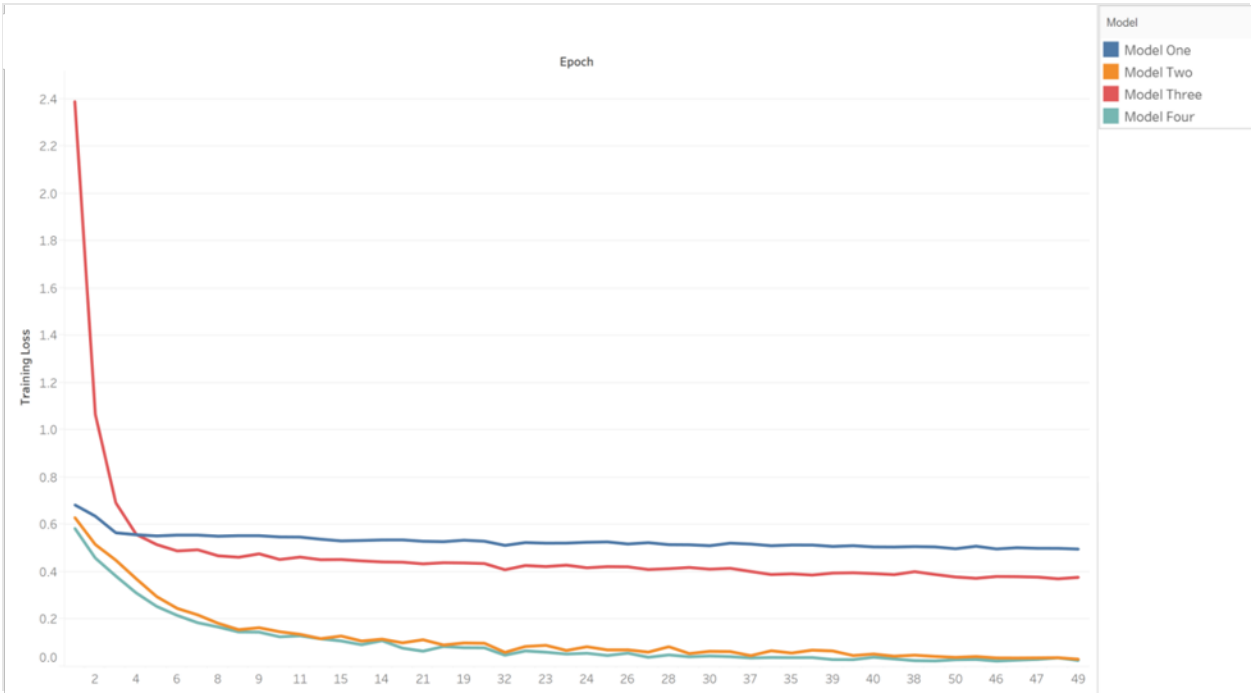
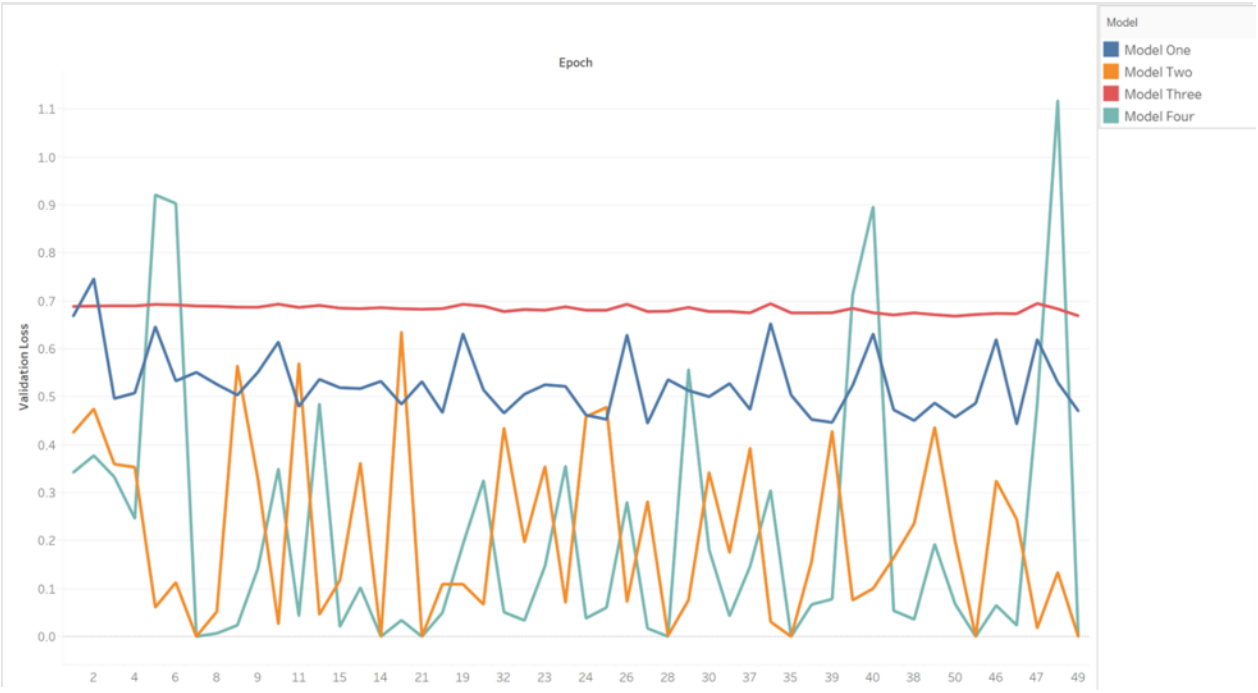


Figure 18: Model Comparison: Validation Losses



Of the four models, Models Two and Four had the best training and validation accuracy results. This was to be expected as Models Two and Four relied on a pre-trained model and were able to utilize transfer learning to achieve better results than the models with simple, not previously trained, convolution neural networks. Between Models Two and Four, Model Four had slightly higher accuracy across both its training and validation datasets. This would suggest that employing an automatic validation split rather than manually splitting the data results in a slightly more accurate model. However, these results are likely not statistically significant. Because the difference in accuracy between these models is small, it does not necessarily mean that one validation approach is better than the other. It is important to remember that, because of the dataset imbalance, accuracy may not be the best metric to understand the classification value of these models and further evaluation is recommended. As can be seen in Figures 11 and 12, some models tend to demonstrate periodicity, or a seemingly repetitive behavior, and this is an algorithm behavior that should be explored in future research. Running the model for more epochs could provide a clearer picture of how the model is behaving. Accuracy metrics of these models should be considered in tandem with other model characteristics, such as the size and type of the dataset.

In addition to the training and validation visualizations and results, it is important to also evaluate the trained models on the test dataset. For purposes of continuity, the research team relied on the use of an evalutate generator. The evaluate generator tests the overall accuracy of a model and provides a single metric to indicate its results. This is a general evaluation to provide a sense of how well a model is learning and classifying, and this function does not allow the research team to determine which images specifically were misclassified. Below, in Table 1, are the evaluation results from all four models. From the evaluation function, it is clear that Model Four had the best results with an accuracy of around 95%.

Table 1: CNS Model Evaluation Accuracy Scores

Model One	0.7385
Model Two	0.9365
Model Three	0.7531
Model Four	0.9531

To further evaluate these models for future similar research, it could be beneficial to save the models and images to a device's memory and perform predictions on single images so as to determine which particular images the model is having difficulty with and any trends across the image that might indicate features with

which the model is specifically struggling. Because the purpose of this report is a proof of concept, this evaluation approach was out of the scope of this research.

Conclusions

Model Assesment

By building these image recognition and classification models, it was the goal of the research team to assess the potential utility in developing classifiers for any and/or all controlled goods that could eventually be deployed alongside personnel implementing strategic trade controls and detecting any potential violations in these controls. Based on this stated goal and respective successful model results, it is clear that a machine learning classification model could be a useful tool to employ. However, the demonstrated limitations must also be considered.

From the results, it is clear that machine learning models are able to accurately and effectively identify controlled dual-use goods. Based on the provided loss metrics, it is evident that the machine learning models would be most efficient in image recognition and classification when employed alongside a human, such as an individual implementing strategic trade controls or assessing potential control violations. Using an image classification machine learning tool, the issue of identifying proliferation-sensitive dual-use goods can begin to be addressed, supplementing nonproliferation efforts through enhanced strategic trade controls. The models outlined here could be further developed to help personnel without specialized knowledge implement strategic trade controls on dual-use goods by serving as an “image classification dictionary” of sorts. This assistance from machine learning would be more advantageous in implementing controls than monitoring efforts that only relied on human regulators.

The limitations outlined in this report are also worth considering when discussing the viability of future machine learning applications. It is important to note that controlled items might have a different context when they are being traded internationally. There may be clutter—or features that distract from the main object of interest—surrounding items of interest depending on how and where they are being shipped that might affect image classification capabilities. With this in mind, perhaps the largest and most important consideration when moving forward with additional research is image dataset size and availability. Despite the accurate model results presented here, there needs to be a wider availability of images to advance research in this field. While limited access to controlled and dual-use goods can benefit

nonproliferation, larger—and more diverse—datasets are necessary for continued model exploration. A potential solution to this issue could be resampling so that all classes are equally represented. Even still, as discussed by the complications surrounding the data-collection process, it is presently difficult to build a dataset that is uniform and representative of all controlled items that would be ideal for image classification. While the research team attempted to avoid biases in the dataset formation, the limited availability of images made it difficult to ensure image diversity. Ultimately, the dataset determines the success of the machine learning model, and this case, must be improved for future field applications.

An important next step in this research would be developing larger datasets of more classes that could be further illustrative of controlled goods and improve the training and validation of machine learning models. Similarly, in this dataset expansion, explicitly selecting additional images for training and validation that would test "edge" cases—where the machine learning algorithm could potentially struggle with visually similar items—would be beneficial. This would require discussions about best practices in building a large dataset for training a machine learning algorithm, including discussions identifying biases that exist in pre-existing datasets and how to best avoid them. Additionally, this would require experts and end-users to identify objects the classifier might be shown that could conceivably be similar enough to cause errors.

Because of the criteria that decide which dual-use goods are subject to export control, it is unlikely that full automation by means of machine learning image classification is likely. While image classification models can recognize objects, often they cannot recognize context nor specific item characteristics such as material and size. The importance of context when assessing the proliferation potential of a dual-use good is not eliminated with machine learning models. As demonstrated in the model evaluation section, it is clear that the models struggle with some images and classes, particularly those with complex and/or similar settings. This problem of context could be a limitation of the dataset and could be further improved upon contingently with dataset improvement. However, even with dataset development, it is likely that a machine learning model will be more accurate in classifying and identifying certain images over others. This non-uniform application, paired with the diversity of goods that fall under strategic trade controls, make it clear that machine learning models alone cannot solve the issues of dual-use strategic trade controls outlined in this report.

Potential Machine-Assisted Applications and Future Research

As briefly described above, it would be useful to apply machine learning classification tools alongside implementers of strategic trade control in order to improve their dual-use object recognition capabilities. As the models currently stand, and with subsequent further development, field applications could take the shape of an image dictionary of sorts. If an implementer of controls does not have specialized knowledge of what a specific item looks like, or whether an item is controlled, they could rely on a machine learning tool to provide them with said knowledge, thereby improving the efficacy of strategic trade controls. Machine learning tools could inform implementers what a dual-use controlled item looks like alone and in an operational context, as well as how the object might be used for WMD proliferation. As research progresses and datasets are improved, additional goods could subsequently be added to this reference database, proving to be more effective in preventing WMD proliferation.

Immediate future research endeavors related to image classification machine learning models to help identify controlled dual-use goods for nonproliferation efforts should focus on expanding and improving relevant, publicly available datasets. To do this, a separate machine learning tool could be employed to monitor relevant websites and automatically tag or classify new images based on the existing trained models and respective datasets. Additionally, a scraper tool could be employed to automatically extract relevant images from videos and various websites. These images could then be used to create improved training and validation datasets.

Finally, it would be useful for continued research to assess the feasibility of field applications alongside regulatory agents and warfighters. It would be of value to test mobile phone applications, or similar field applications, to see if it provides machine-assisted detection of dual-use, controlled items.

The research outlined in this report is an important step in demonstrating the potential value of machine learning tools can have to solve in solving some of challenges of strategic trade control applications. While more research and modeling are needed, in the future, these image classification tools can assist regulators and improve the efficacy of strategic trade controls for nonproliferation efforts.

About the Author

Jamie Withorne is a research assistant at the James Martin Center for Nonproliferation Studies (CNS), an affiliate of the Middlebury Institute for International Studies at Monterey. At the CNS Washington DC office, Withorne conducts data analysis on a variety of topics, including machine learning, quantum technology, sanction evasion tactics and trends, and the nuclear fuel cycle. Withorne has previous experience at Columbia University's School of International and Public Affairs, the US Department of State, the Center for Arms Control and Nonproliferation, and the American Enterprise Institute. Withorne holds a BA in Political Science from Columbia University in the City of New York.

About the Nonpro Notes Series

The Nonpro Notes series is intended for the CNS DC team to share specific insights resulting from our research. The series generally focuses on methodologies, tools, and data produced from our work that is more narrow and specific in scope than would generally be suitable for publication in a peer-reviewed journal. At the same time, it is envisioned that the insights shared through this series can be built upon in future research and publications.

www.nonproliferation.org/dc