Analysis and survey of Auto Machine Learning (AutoML) for Computer Vision-based Gender Recognition

---

A Capstone
Presented to
the Graduate School of
The University of North Carolina Wilmington

In Partial Fulfillment
of the Requirements for the Degree
Masters of Science in Computer Science and Information Systems
Computer Science

---

by
Ethan Tyler Cook
Dec 2022

Proposed to:
Dr. Karl Ricanek, Committee Chair
Dr. Geoffrey Stoker
Dr. Hosam Alamleh

Abstract

The market for artificial intelligence is one of the fastest-growing markets in the World. The market value was estimated to be approximately $65.48 billion in 2020 and is expected to grow 38% before 2030.[Keshav K, 2022] Deep learning is a subcategory of artificial intelligence (AI) that is contemporary cutting-edge technology. Deep learning is a subclass of machine learning algorithms that uses multiple layers to progressively extract higher-level features from raw data. There are seven steps in building a deep learning model include: gathering data, preparing the data, choosing the model, training the model, evaluating, hyperparameter tuning, and predicting. Understanding each step and how to build a good-performing model takes years of experience. That is where AutoML (automatic machine learning) can give the power to build one of these complex models to laymen and novices alike. AutoML automates every step in building a model besides gathering data. With basic knowledge of computers, anyone can train and deploy a complex predictive model. This will not only give novices the power of AI, but it will also free up the employees that are tasked with building and training models that could just be solved with AutoML. This will allow data scientists to focus more time and resources to the more complex problems. Companies that thrive in this domain can anticipate billions in revenue. In my paper "Analysis and Survey of Auto Machine Learning (AutoML) for Computer Vision-based Gender Recognition", I will research how AutoML works on a common but challenging problem in the computer vision space. Further I want to investigate if AutoML can deal with unbalanced data without creating biased predictions. With the growing problem of algorithm bias, AutoML will have to mitigate or remove these biases to make fair predictions. Algorithm bias is systematic and repeatable errors

in a computer system that create unfair outcomes, such as privileging one arbitrary group, or class of data, of users over others. The questions that I want to answer are: How does AutoML compare with hand-tuned deep learning models in terms of performance? How sensitive to algorithm bias is AutoML? Is AutoML a solution to reduce or remove algorithm bias due to systemic racism?

Table of Contents

Title Page . . . . . **i**

Abstract . . . . . **ii**

1 Introduction **1**
- 1.0.1 AutoML 1
- 1.0.2 Gender Classification 4
- 1.0.3 Algorithm Bias 4
- 1.0.4 Research Questions 6

2 Background **7**

3 Methodology **15**
- 3.0.1 Microsoft Azure System 18
- 3.0.2 Measurements 22

4 Experiments **26**
- 4.0.1 Experiment 1: 28
- 4.0.2 Experiment 2: 28
- 4.0.3 Experiment 3: 29
- 4.0.4 Experiment 4: 30

5 Results **38**
- 5.0.1 Experiment 1 Results 38
- 5.0.2 Experiment 2 Results 39
- 5.0.3 Experiment 3 Results 40
- 5.0.4 Experiment 4 Results 40

6 Conclusion **43**

Bibliography **46**

Chapter 1

Introduction

1.0.1 AutoML

AutoML is short for automated machine learning. The entire process of building a data pipeline is automated from data processing and cleaning to feature extraction to model evaluation. The only element that is not automated is data collection. With enough data, anyone with basic computer knowledge can build a deep-learning model. With this technology it will give companies with smaller budgets the ability to use the data they have collected to make references or help with predictions.

AutoML has been a hot topic for many years now and for good reason. AutoML started in 1995 and the only part that was automated was the parameter optimization step which is to fine-tune the selected model. This was a big step in artificial intelligence history because it allowed for quicker and more efficient tuning of models. Then it branched into a model selection step which is where the best model is selected for the use case. This is where the NAS (neural architecture search) is implemented. NAS helps build and train custom deep learning architectures. There are many different types of NAS that yield a variety of results for different problems. Thanks to the NAS, technology advancements in model selection have allowed

AutoML products to select multiple models. It will train all the child models to a certain point. From that point on, the best model is selected for further training. NAS started out being very computational expensive taking 800 GPUs over 2 weeks to complete.[Thomas Elsken, 2019] Now it takes Azure's Custom Vision AutoML product a couple of hours on a cloud service to train and evaluate a model. Since 2010, AutoML has completely automated the process of building a predictive model. The process has transitioned from data cleaning to model evaluation. Data collection is the only part of a data pipeline that currently, is not automated. However, the future may show that this step can also be automated given the vast amount of data that proliferate the world today.

This is important because AutoML is giving the power of artificial intelligence (AI) to novices in many different fields of work. Building an AI model takes many years of experience and it is a trial-and-error process. This means it takes experience, time, and money to become proficient in designing and building a predictive model. AutoML handles each of these problems. With the Azure Custom Vision AutoML product it makes it quick and easy to build a predictive model. Also, with this product being a part of the cloud services it is relatively affordable. AutoML will help increase efficiency, cost-savings, accessibility, and performance for companies and individuals that implement this technology. Efficiency is increased because it simplifies and speeds up the process of building and training a model. By default you are saving money and resources for the company when you can increase the speed of a process. It helps with accessibility because it simplifies the building and training process, making it easier to train staff on building predictive models. Also, freelancers will be able to offer artificial intelligence without taking a data science class. Lastly, it increases performance because in some cases, the models built by AutoML have been proven to be more efficient than hand-built models. [Thomas Elsken, 2019]

This will help companies that are behind in AI technology, because it will free up their data scientists or allow them to build and maintain a predictive model in house instead of outsourcing work. AutoML will free up data scientists by solving some of the more trivial yet time-consuming problems. For example, some age predictive models have different pipelines for different genders. AutoML can build gender classification models that can be trained in less than an hour and get accuracy in the high 90s with only a couple hundred images in the dataset. This is something I tested in the beginning of my research by using images from the internet. Before hand, this would taken an expert data scientist multiple days to architect a model and train it on thousands of images. Also, the data scientists would have to clean the data and perform some feature extraction on the images which is automated in AutoML. Now this task of gender classification can be handed down to a less experienced engineer. This will free up the data scientist to focus their work on more complex problems. This will allow companies to speed up and scale up their AI solutions. AutoML is on the rise and is being tested by many different companies. These companies include big names like Google's Cloud AutoML, Amazon's SageMaker Autopilot, DataRobot, and Azure Automated machine learning services but there are also open-source solutions like Auto-Sklearn and H20AutoML. This technology will be a big part of AI in the future and isn't going anywhere. The companies that prevail in this space will amount to millions if not billions of dollars in revenue.

AutoML is currently being studied and used by many different professionals in different fields of work. There is a lot of research on using AutoML in the medical field to help diagnose certain diseases and cancerous cells. For example, there was a study done using Google's Cloud AutoML Vision to detect invasive ductal carcinoma. In this study, they were able to get a 91.6% accuracy for their model evaluation and an 84.6% accuracy on a held-out dataset. [Yan Zeng, 2020] This would help the

professional recognize and diagnose these patients in the early stages. Professionals that have data that they have collected, would be able to use that data to help build an AutoML predictive model. That model could then be used to draw insights or highlight areas to focus on so the professional would have help evaluating all the data.

1.0.2 Gender Classification

Gender classification is a problem where the model is given an image of a person and the model needs to predict the gender. These problems are known as binary classifiers since there are only two options to pick from. This kind of problem is being used to help design more complex models to guess a person's age from an image. Some of the models are used to guess chronological age, so first, they want to separate the males from females to run them through different pipelines. This is because some models are believed to guess a more accurate age estimation if you train separate models for males and females. [Guodong Guo, 2009] There are also products that solely do just that. They take an image and give a prediction of what gender the person is. Gender classification has also been a hot topic lately due to its vulnerability to algorithm bias.

1.0.3 Algorithm Bias

"Algorithmic bias describes systematic and repeatable errors in a computer system that create unfair outcomes, such as privileging one arbitrary group of users over others".[Florida State University, 2016] Some algorithmic bias can come from under-sampling for certain subgroups, and classes, within a dataset but sometimes this bias is prevalent in models trained on abundant and balanced datasets. In a study done by Vítor Albiero, they found that algorithm bias can also come from the fact that women wear make-up and have less facial information due to more head hair or small

faces.[Vítor Albiero, 2021] Algorithm bias can come from many different problems but it's key to try and remove or reduce as much as possible to make a fair prediction. So overall accuracies for a model are not a fair analysis of the performance. The overall accuracies need to be broken into accuracies for each group of the dataset. This will show if an algorithm is biased toward certain subgroups. To get the best and fairest prediction each subgroup should have close to the same accuracies. The paper Gender Shades by Joy Buolamwini and Timnit Gebru [Buolamwini and Gebru, 2018] tested well know facial analysis products to look for gender and skin-type biases. In the experiment, they trained these products with a balanced dataset for each subgroup in their data. It was found that females and darker-skin-toned individuals had the lowest accuracies. The face analysis product for IBM classified white males with an accuracy of 99.7% but classified darker females with an accuracy of 65.3%. That is a 34.4% difference which shows a strong bias toward white males. When broken down more it was found that females with the darkest Fitzpatrick Skin Type were only accurately classified 46.8% of the time. With that low of an accuracy, the model is pretty much doing a coin flip. An accuracy that low would be close to implementing a random guesser with the theory that 50% of the time it will be correct.

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

Figure 1.1: Here you can see the break down of each subcategory on all three facial recognition products based on accuracy. It can be seen that the darker females seem to the worst overall for each product. Also the difference between male and female accuracy shows how males are predicted with a lower error.
[Buolamwini, 2018]

This is a problem because these biases could be passed down through the pipeline. This is a serious problem for facial recognition and facial analysis, and further for all deep learning solutions. Facial recognition is being used by our law enforcement to help catch and identify suspects. This means that darker individuals would be more likely to be wrongfully identified and wrongfully convicted. Facial analysis is being used by companies to help with who gets promoted, who gets a raise, and who gets hired or fired. So, biases in these domains will cause a major reset on all the advances we have made in our society to create equality for everyone. [Buolamwini and Gebru, 2018]

1.0.4 Research Questions

Using Azure's Custom Vision, I will address the following four questions:

- How mature is AutoML to handle computer vision tasks like gender classification?
- Can if anyone with basic computer knowledge with no machine learning experience can build a predictive model with AutoML?
- How AutoML compares to traditional hand-built deep learning models?
- How sensitive AutoML is to algorithm bias. Also, whether AutoML is a solution to reduce or remove algorithm bias, or will it propagate the problem?

AutoML is designed to create a top performing solution for the untrained and therefore, could inadvertently build solutions that is biased. AutoML has the potential to become a well-known product for companies. if AutoML does not compare well to hand-built models or has significant algorithm bias, then it will never become a viable option.

Chapter 2

Background

Companies are becoming more interested in the technology of AutoML from the manufacturing to consumer applications, AutoML has been a popular topic in recent headlines. One of the reasons for this popularity growth is the advancements in neural architecture search (NAS) algorithms. Also, the increase in computation power brought on by cloud computing leverages powerful hardware–to include large pools of GPU's–that cuts down on designing and training times. With the raise of the AI Economy, the growth of NAS algorithms, the use of open neural network exchange (ONNX) an organic growth in AutoML is occurring. The question that this work is investigating is how prepared are the current technologies as inferenced by the Microsoft Azure platform. Does the Azure AutoML platform generate bias for its solution?

In a study using Google Clouds AutoML product, researchers tested the feasibility to tackle the problem of identifying invasive ductal carcinoma (IDC) using whole slide images (WSI). Invasive ductal carcinoma accounts for 70-80% of all breast cancer diagnoses.[Yan Zeng, 2020] For their experiment, the researchers wanted to use AutoML to help interpret and diagnose invasive ductal carcinoma from different levels of whole slide samples. This task is usually done by a pathologist that has to

analyze large areas of whole slides and different levels of magnification. Also, this task is done in a clinical setting that demands these analysis to be completed accurately and quickly. Google's Cloud AutoML Vision was able to architect, build, and train a model with an accuracy of 84.9% on a holdout dataset. The researcher compared their results to earlier works that used CNN's and Deep Neural Networks. They found that their accuracy was the highest out of the group proving that Googles Cloud AutoML Vision is a feasible solution for identifying invasive ductal carcinoma.

Other researchers have built and evaluated other solutions to this problem. The solutions that were used for comparison in Yan Zeng and Jinmiao Zhangt [Yan Zeng, 2020] research was a CNN that was built during the research of Cruz-Roa's [Angel Cruz-Roa, 2014] mentioned in their paper and Alexnet which was evaluated in Janowczyk's paper. The AutoML product they tested used was Google Cloud Vision which is Google's cloud machine learning service for image classification. In their research, they found that Google Cloud Vision was a mature and feasible solution for their problem. This was due to the fact that their solution outperformed both the CNN and Alexnet on this problem domain. The accuracies for the CNN and Alexnet were 84.2% and 84.6% respectively. The model that was built by Google Cloud Vision had an accuracy of 85.2% for the model evaluation and 84.6% on the balanced hold-out dataset. Google's Cloud Vision not only completed this task without assets but was able to outperform the prior solutions. This paper demonstrated that AutoML can tackle a challenging computer vision problem. With using a Google product they were also able to take full advantage of using cloud products. With flexible deployment options to elastic infrastructure this would make it easier for a novice to learn and deploy a live instance of their model. That could create problems if the service costs are too much for novice to justify but Google's prices seems reasonable as seen in figure 2.1.

The problem that I will be using to evaluate AutoML will be gender classifica-

Cost for AutoML vision model.

Component	Type	GCP Charge Rate	Resources for Experimental Model	Cost for Experimental Model
Cloud storage	Standard bucket	$0.026/GB/month	3.8 GB	$0.10/month
Model training	Binary classification	$3.15/node hour	4.4 node hours	$13.86
Model deployment	Batch serving	$2.02/node hour	1 node hour	$2.02
	Online serving	$1.25/node hour	1 node, 6 clock hours	$7.50
	Offline serving	Free	<1 GB model files	Free

Figure 2.1: Here is the cost break down for using Google Cloud AutoML with other cloud options [Yan Zeng, 2020]

tion problem. Gender classification is the problem domain that can be described by classifying a face as either male or female. This type of problem is a binary classifier problem. This problem has become a popular topic due to its link to algorithm bias in face-processing algorithms.

I will compare the performance of the AutoML system of Microsoft with a hand-tuned solution. I have chosen to use the work of Philip Smith [Philip, 2020]. In the research done by Philip Smith [Philip, 2020], he proposed three methods for mitigating algorithm bias within predictions. When mitigating any bias that may occur in a certain subcategory of your data without affecting the predictions on other categories then subsequent leads to higher accuracies overall. The techniques that were used to mitigate bias were data augmentation, dataset oversampling, and synthetically generated data. Data augmentation takes the training data and applies different filters to the image to manipulate it. These filters could change the color of the image, translate the image, cut out parts of the image, and that's just a few examples of filters that can be used. Figure 2.2 shows what the data would look like after applying some of these filters. Dataset oversampling takes the subset of data that is being predicted at a higher error rate and uses more examples of that subset to force the network to learn more features from the harder examples, i.e. the sub-groups with the worst performance. Figure 2.3 shows the percentage breakdown of his oversampling technique. This will give the model more data points on that

subset, giving the model a better chance of distinguishing that subgroup. Last, is synthetically generating data which uses a generative adversarial network (GAN) to synthetically generate data for underrepresented areas of the dataset. Figure 2.4 shows a synthetically generated image from a GAN where the image of a woman is aged from 25 to 55 years old. According to Philip this process can augment a dataset or sub-group that has less number of images. It can also be used to augment the overall dataset size, growing it by factors.

Figure 2.2: This figure shows what the data would look like after being augmented with his data augmentation policy.
[Philip, 2020]

They were able to test each method and combine their finds to create a predictive model with higher accuracies. In this research, Smith was able to propose a convolutional neural network (CNN) with a state-of-the-art performances. The performance (accuracy) that was reported was 99.82% which is the state-of-the-art for

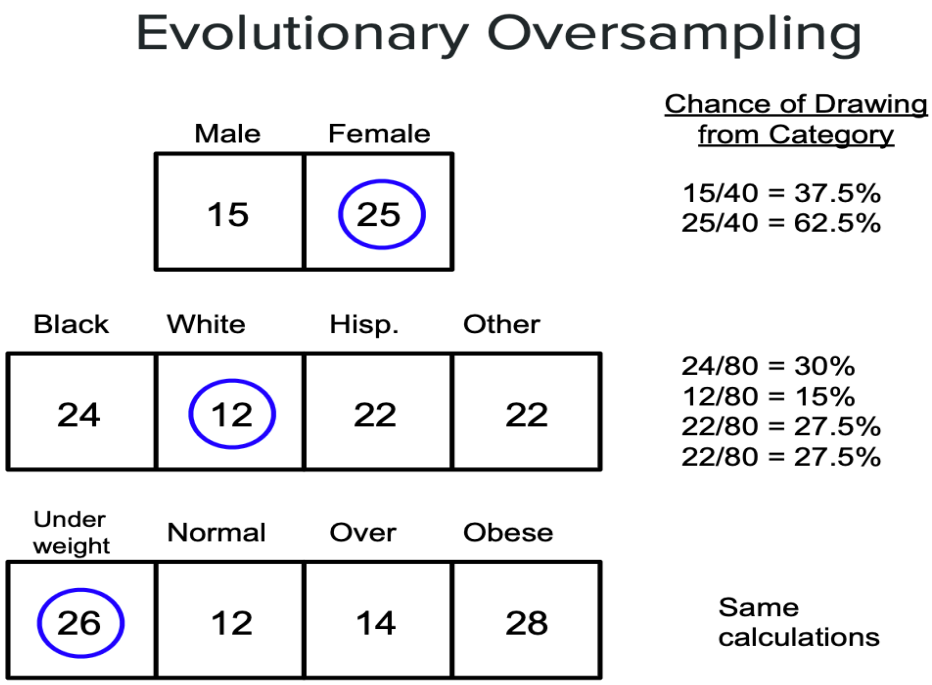


Figure 2.3: This figure shows the oversampling breakdown proposed by Philip. [Philip, 2020]

Figure 2.4: The 25 year old woman in the image is aged using a GAN to produce an image of what she would look like at the age of 55. [Philip, 2020]

this problem currently. This performance will be used as reference to compare my studies to.

In Philip Smith's [Philip, 2020] research he was trying to mitigate or remove algorithm bias from his predictive model. Florida States' definition of algorithm bias is described as "systematic and repeatable errors in a computer system that create unfair outcomes, such as privileging one arbitrary group of users over others. Also, occurs when an algorithm produces results that are systemically prejudiced due to erroneous assumptions in the machine learning process". [Florida State University, 2016] Algorithm bias is becoming a big topic within artificial intelligence because of hidden harm that could cause setbacks in our society. This is a growing concern when the stakes of the tasks to be performed are at the greatest expense to people. An example of algorithm bias and its hidden harm can be seen in Cathy O'Neil's book Weapons of math destruction: How big data increases inequality and threatens democracy.[O'Neil, 2016]

In the book, she discusses algorithms that help with calculating a probability map to predict where crime hot spots are. These models do not use racial, religious, or gender data but they do use zip codes, crime levels, and low education levels statistics. This can lead to heavier policing in areas that have lower incomes and lower levels of education. Since people with lower incomes and lower educations levels tend to concentrate in certain zip codes this leads to a cycle of bias and discrimination. Just because someone lives in a low income area with lower education does not mean they are a criminal. This books shows that algorithms with bias do not harm everyone equally but they focus in on the poor, marginalized, and vulnerable communities. [Woodson, 2018]

This algorithm has the ability to make recommendations for sentencing. The research demonstrated that sentencing was extremely biased; ethnic minorities faced

stiffer sentencing by the AI for less serious crimes then ethnic majority. The output from an algorithm that has a underlining bias can feed that bias down the line causing biases throughout the entire data pipeline. Meaning one part of an networks architecture could cause many problems throughout model, giving inaccurate predictions or insights.

In the research done by Joy Buolamwini [Buolamwini and Gebru, 2018], she looked at 3 commercial gender classification products to understand how they perform on certain subcategories. She noticed that these commercial products give an overall accuracy but not accuracy on each subcategory. Also, she noticed that being an African American female that the products had a hard time predicting her gender when compared to her more fair-skinned colleagues. In her research, it was found that all three products had increased error percentages with darker individuals and females. One product that was tested is the IBM gender classifier which had an overall accuracy of 87.9% but an accuracy of 65.3% on darker females. With these findings IBM conducted their own study to address this problem of algorithm bias.[IBM, 2018]

More research needs to be done in the area of algorithm bias to help protect our futures from an algorithms prediction that could ensue a prejudiced system. Without demanding transparently for all subcategories results, biases will go on to live in these high-stake systems forever. Algorithm bias could negatively impact the "civil rights movement and women's movement under the false assumption of machine neutrality. We must demand increased transparency and accountability." [Buolamwini and Gebru, 2018] Figure 2.5 shows some of the potential harms that can arise from algorithm bias.

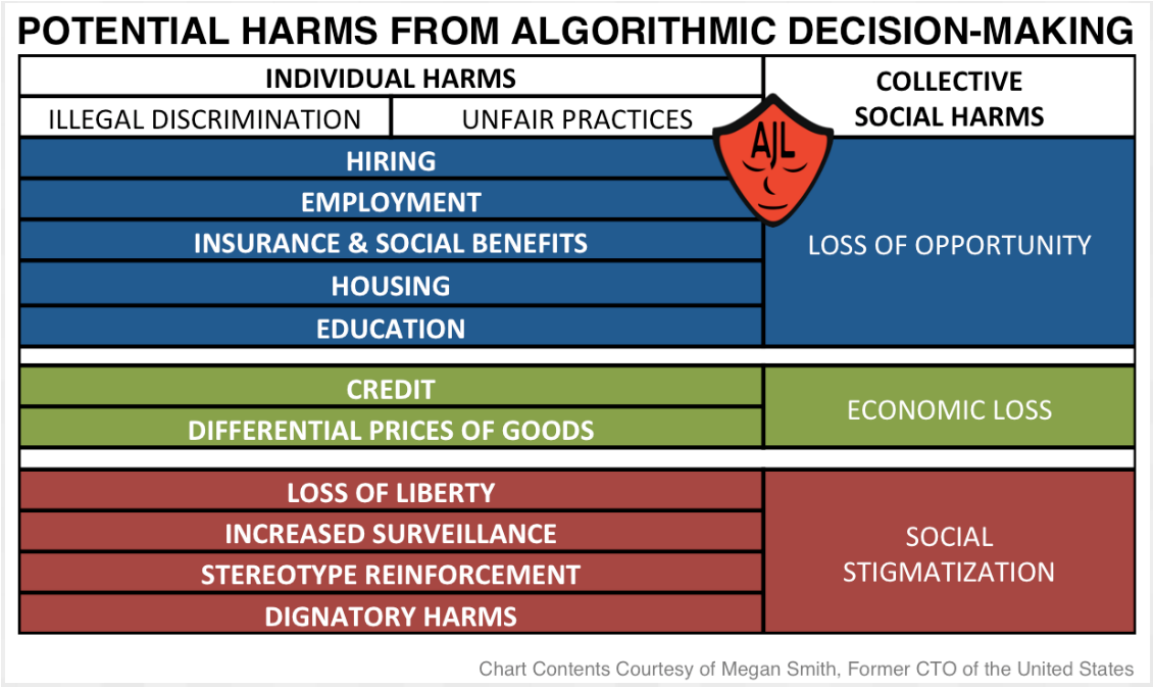


Figure 2.5: This figure shows some of the potential harms that can come from algorithm bias. It can be seen that algorithm bias can and will affect many different parts and peoples lives. [Buolamwini, 2018]

Chapter 3

Methodology

The process of designing and building predictive models has been a task that only a small number of professionals have the knowledge to perform. With AutoML this process can then be performed by anyone with basic knowledge of computers. Further, AutoML does not require a large amount of data to design and build predictive models. Azure's Custom Vision AutoML product claims that it can build a predictive model with as few as 50 images per each category. AutoML will change the way that we think and plan artificial intelligence tasks, but the question is how mature is AutoML. If AutoML makes it easy to help deploy predictive models in applications but has bad results, then *AutoML may be at risk for causing significant harm to our society. AutoML solutions must have comparable results to hand-built models and it must not exhibit algorithm bias or differential performance.* Without meeting these basic objectives, AutoML solutions will not be accepted by society or governments.

With the growth of artificial intelligence, there's a push for more clarity on the results of all subgroups. So, not only does AutoML need to be comparable to hand-built models but needs to design ways to limit algorithm bias from propagating throughout the predictive model. In my research, I hope to answer both of these

questions using Azure's Custom Vision AutoML product.

The approach taken in my research is to compare Azure's Custom Vision to the work of Smith [Philip, 2020] which contained top performance for gender determination on the dataset that is being used in this work. In Smith's research, he was trying to mitigate algorithm bias using the tasks of gender classification, age estimation, and body mass index (BMI) estimation using only an image. I will be using his results from the gender classification to compare to my results from Azure Custom Vision. In his research, he was able to build a gender classifier that had state-of-the-art performance numbers for gender classification which produced an accuracy of 99.82%.

I will be using the MORPH dataset which is a collection of inmate photographs. The MORPHs dataset has images that all have close to the same lighting and distance between the camera and the subject. This dataset was also used in Phil's research. The dataset is made up of a total of 55,608 images. Since the dataset is made up of photographs of the same individual over a course of time, I will only using one photograph of the person. This takes the number of images that are usable to 13,673 images. The dataset has more males than females and more African Americans than white individuals. There are mostly African American males which make up 67.7% of the dataset. The breakdown of the dataset can be seen in figure 3.1.

To test AutoML to its full potential, I will have multiple experiments that have different size training datasets. In total, I will perform 4 main experiments with 3 sub experiments for a total of 12 experiments. The first experiment will have the least amount of training data. The same training data will be used for building three different models. Each model will have a different training time assigned to it.

The first will be using the quick train option which stops training once a threshold is met. I will record the time that each quick training session runs. The

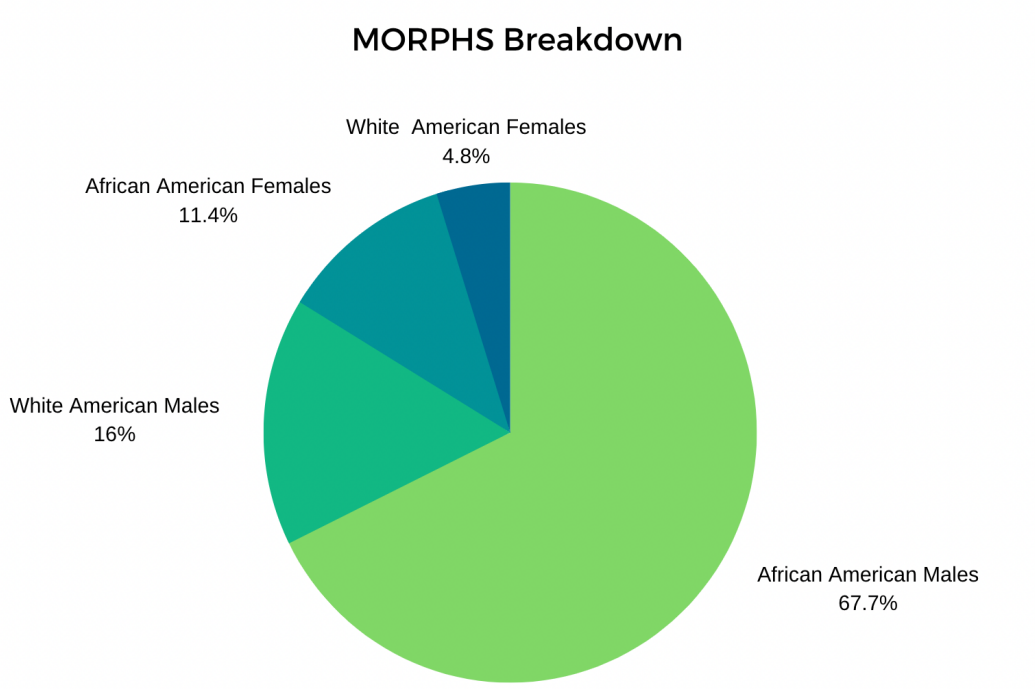


Figure 3.1: The MORPH dataset is made up of 55,608 images of 13,673 different inmates. The dataset is mostly African American and mostly male. African American males make up 64.8%

next experiment will be using a 2-hour training limit. Then the last experiment will use a training limit of 10 hours. I will build three models for each of the 4 main experiments with different amounts of training data for each main experiment. The first three experiments will be using a balanced dataset meaning that each subgroup has the same number of images in it. This is where I want to evaluate the performance are as a function for training data size. The last set of experiments will be using an uneven dataset. Throughout these experiments, I will be testing for algorithm bias. If AutoML can handle being given uneven data to train itself with and not create a biased decision, then it might be a solution to reduce or remove algorithm bias. In each experiment, I will be analyzing the results to compare them to the hand-built model and to look for any algorithm bias. Figure 3.2 has a table that itemizes all the

experiments for convenience.

Exp #	Time	Training Data Size	Training Data Characteristics
1A	Quick train	50	Balanced
1B	2 hrs	50	Balanced
1C	10 hrs	50	Balanced
2A	Quick train	200	Balanced
2B	2 hrs	200	Balanced
2C	10 hrs	200	Balanced
3A	Quick train	500	Balanced
3B	2 hrs	500	Balanced
3C	10 hrs	500	Balanced
4A	Quick train	WAM=450, WAF=150, AAM=150, AAF=50	Unbalanced
4B	2 hrs	WAM=450, WAF=150, AAM=150, AAF=50	Unbalanced
4C	10 hrs	WAM=450, WAF=150, AAM=150, AAF=50	Unbalanced
			WAM = White American Male
			WAF = White American Female
			AAM = African American Male
			AAF = African American Female

Figure 3.2: This table shows each experiments name, training time, training data size, and training data characteristics.

3.0.1 Microsoft Azure System

The product that I will be using is Azure's Custom Vision which is their solution for image classification problems. Azure has created a simple interface to train and deploy its AutoML solutions. When creating a new project in Custom Vision it will ask for a name for the project, a resource (your subscription), the type of project, the classification type, and the domains. The type of project relates to the two options of classification or object detection. Since I'm classifying the whole image and not an object inside the image, I'll only be using the classification option. The classification types are multi-class and multi-label which correlate to a single tag

per image or multiple tags per image. I'll be using only the multi-class option which allows for only one tag per prediction. The last step is to pick a domain. I went with the general option because of my problem not fitting in any other domain. If the problem domain that is being worked on fits in one of the predefined domains these options are better at those select problems. For instance the food domain would be best when trying to classify images of food or detecting certain food in images. Figure 3.3 shows how it looks when you first create a project in Azure Custom Vision.

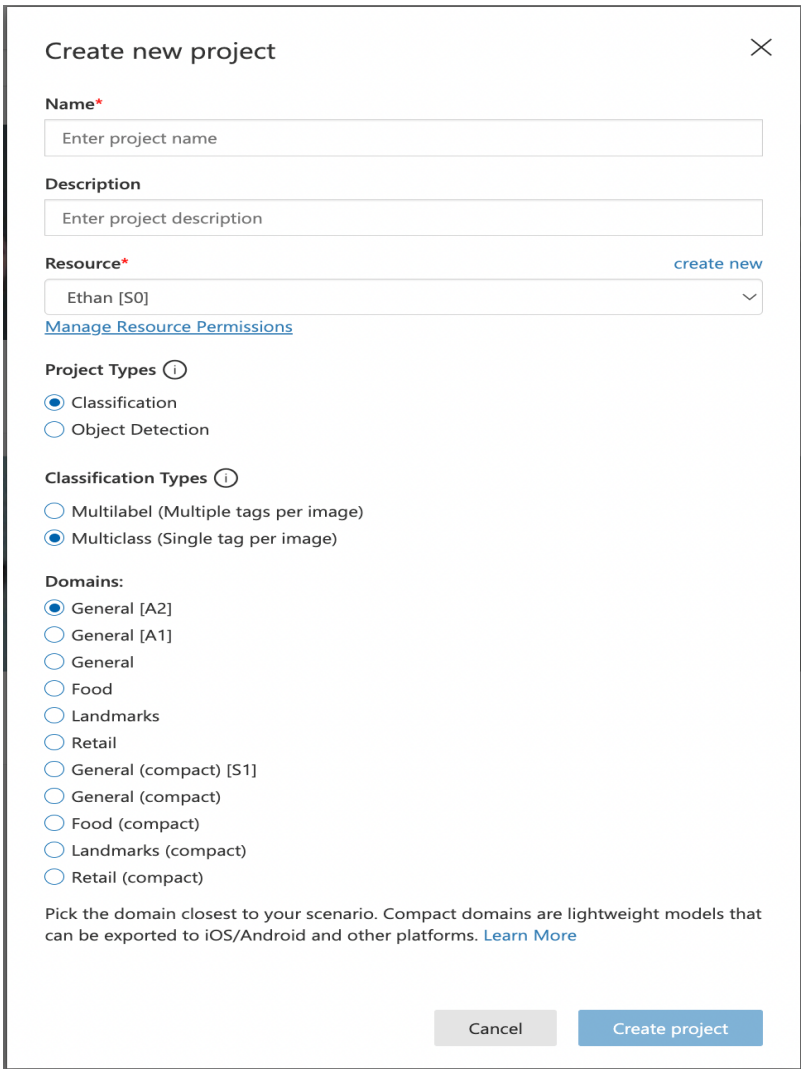


Figure 3.3: This is the first step in building a model with Azure Custom Vision. After this step the project will be created.
[Philip, 2020]

Once the project is created the next step is to load in the data and label them by clicking the upload image button as seen in figure 3.4. Figure 3.5 shows what it looks like when uploading images. During this step you can select or create a new label for the data. After uploading and labeling the data the next step is to start training. This is done by clicking the green train button at the top right of the screen as seen in figure 3.6.

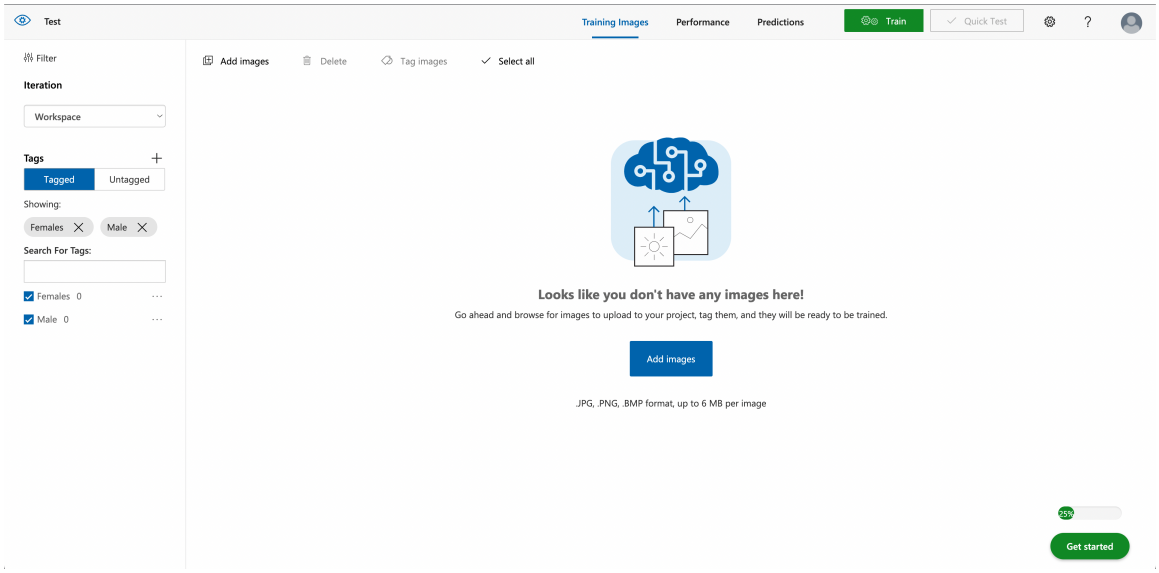


Figure 3.4: During this step the data that is being used will be uploaded and labeled by selecting the button in the middle of the screen.
[Philip, 2020]

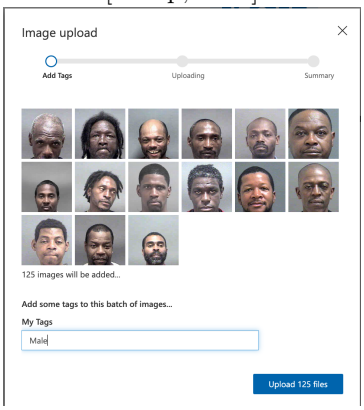


Figure 3.5: Here you can pick from positive or negative label types and also label a batch of images.
[Philip, 2020]

After clicking the green train button it will prompt the user to pick a training

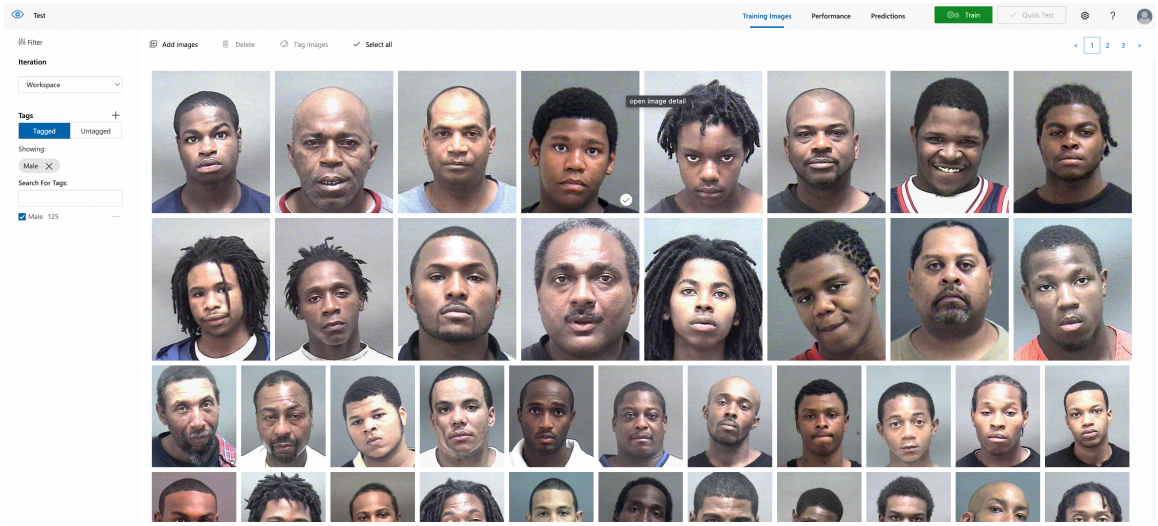


Figure 3.6: Now it's time to train the model. In the top right there is a green button labeled train. Click to select training options.
[Philip, 2020]

type which is either quick train or advanced train. The quick train option will start training right away. The advanced train will ask for a training budget between 1 hour and 96 hours. Both options can be seen in figure 3.7 and 3.8.

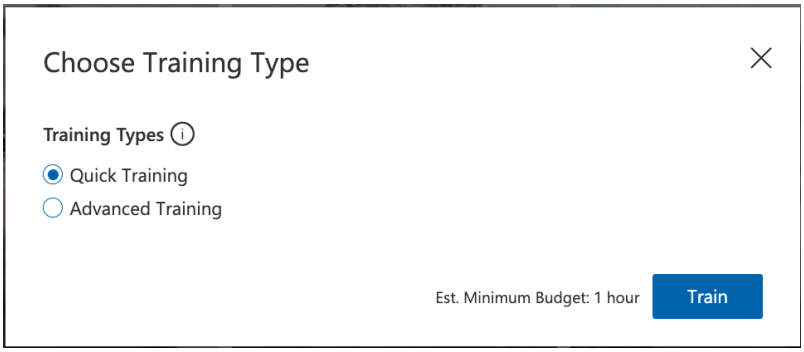


Figure 3.7: Here you can see the quick train option.
[Philip, 2020]

Once the model has been trained, the performance metrics will be shown as seen in figure 3.9. A flow graph showing the process of building a model with Azure Custom Vision can be seen in figure 3.13. Once that is completed the model can be tested right away in the application but only with one image at a time. To run a

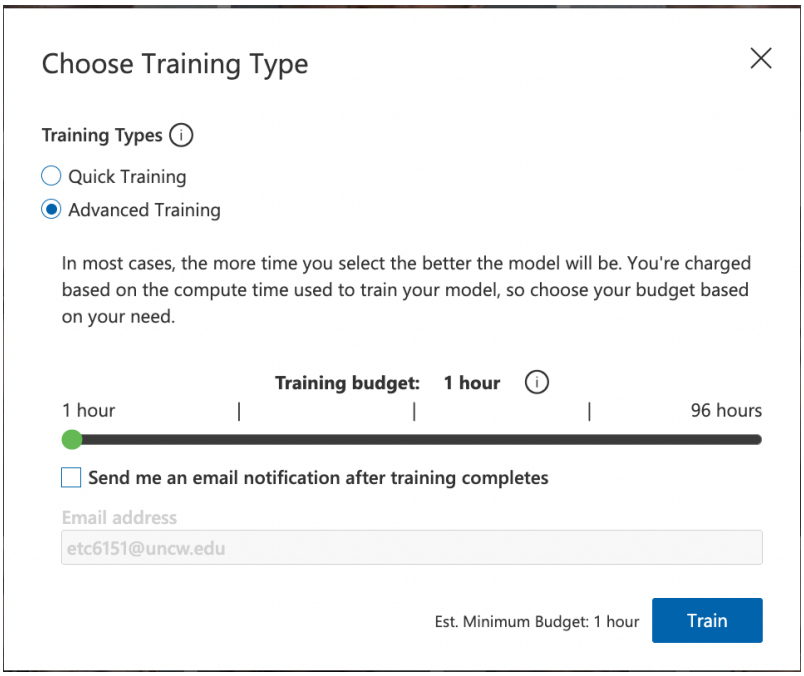
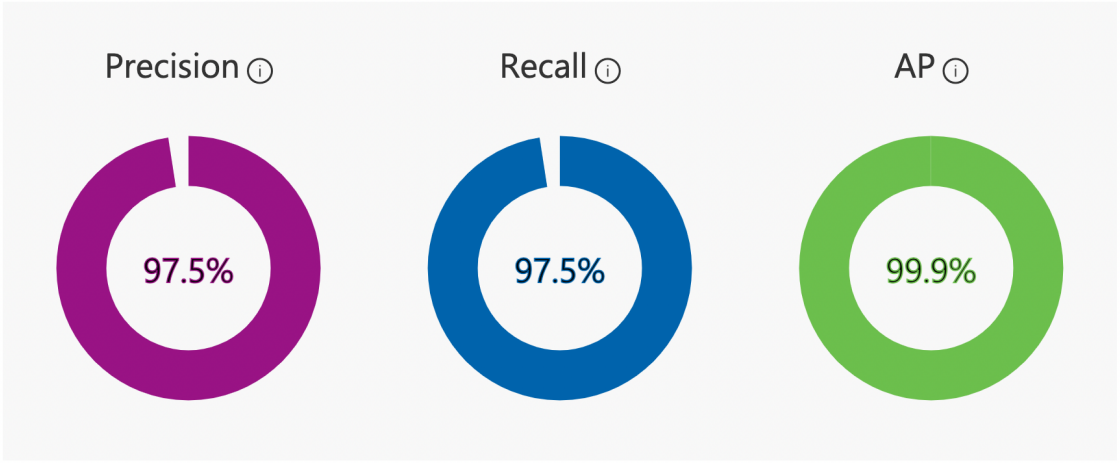


Figure 3.8: Here you can see the advance training option with the training budget slider. This is where you will tell the model how long to train for.
[Philip, 2020]

batch of images, the model must be deployed. To deploy the model that was created in Custom Vision the model must be published. After the model is published it will give a URL and a prediction key which will both be needed to call the model. With the URL and prediction key, a python script can be executed to run a batch of images for predictions. See figure 3.10, a sample of the inference prediction outputs produced by running a batch of images through the model using an API.

3.0.2 Measurements

The performance metrics used by Azure Custom Vision are precision, recall, and AP. Precision is calculated by dividing the true positives by the total number of positive predictions. Precision helps with telling how reliable the model is classifying. Recall is used to understand how sensitive your model is. Precision and recall are used together to drive training and their formulas can be seen in figure 3.11. The last



Performance Per Tag

Tag	Precision	Recall	A.P.	Image count
female	100.0%	95.0%	100.0%	100
male	95.2%	100.0%	100.0%	100

Figure 3.9: The data shown is the output that is generated when calling the Azure Custom Vision API.

id	project	iteration	created	1_probability	1_tagId	1_tagName	2_probability	2_tagId	2_tagName
21b73b27-4a67-4ffd-8bfb-27292e2b2cc2	60eb953d-531b-4d54-b219-2420be806ab1	ae179176-e453-4be4-9dec-5b12b7e919b9	2022-09-08T02:05:36.586Z	0.5000416	1d3963fe-9743-49fa-a872-93a9068c6e08	Male	0.49995837	478583b3-77d9-4314-8c54-6f33e4f1b7d5	female
01b1cd64-6868-4e1c-95af-1d11f4ab569d	60eb953d-531b-4d54-b219-2420be806ab1	ae179176-e453-4be4-9dec-5b12b7e919b9	2022-09-08T02:05:37.112Z	0.9729643	478583b3-77d9-4314-8c54-6f33e4f1b7d5	female	0.02703569	1d3963fe-9743-49fa-a872-93a9068c6e08	Male

Figure 3.10: The data shown is the output that is generated when calling the Azure Custom Vision API.

metric that is used by Azure is an AP score. An AP score or average precision is a measurement of the model's performance which is done by summarizing the precision

and recall at different thresholds. It can also be defined as the area under the curve of a precision-recall curve as seen in figure 3.12.

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

Figure 3.11: Precision and recall formulas
[Liu, 2018]

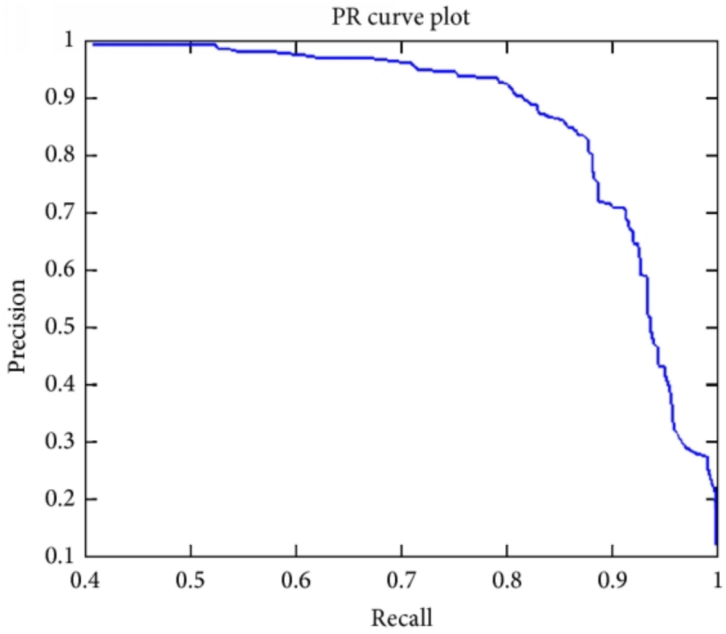


Figure 3.12: The area under the curve is what is called the AP score.

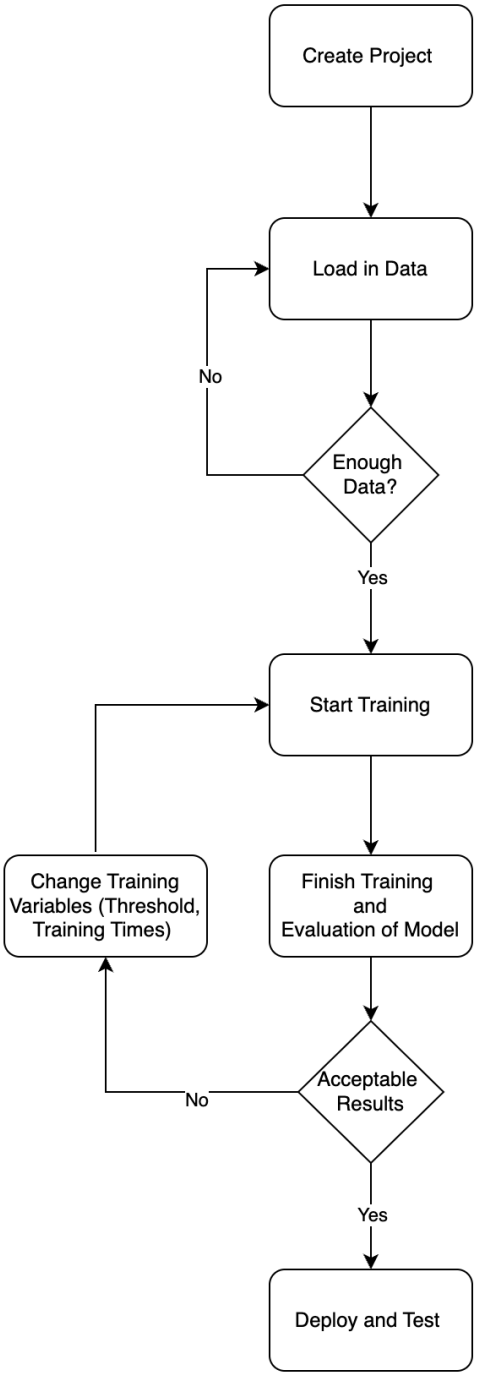


Figure 3.13: This flow graph shows the process of creating, building, and tuning an prediction model with Azure Custom Vision product.

Chapter 4

Experiments

All the data that was used comes from the MORPH dataset [Ricanek, 2006]. This dataset is ideal for gender classification because the images were taken in a consistent way. All the images have close to the same lighting and distance between the camera and the person. This allows for consistent data to train the models with. The data had to be broken up in a way that did not allow duplicate images of the same subject in the same training data. The training data could not have the same images as the holdout dataset.

The holdout dataset consisted of 125 images per subcategory. The same holdout dataset is used as the test set for each of the 12 experiments. This dataset is composed of unique data, e.g. facial images, which was not seen during training. Figure 4.1 and 4.2 shows what some of the data looks like in the holdout dataset. The holdout dataset is broken up into each of the subcategories (African American males, White American males, African American females, White American females) so that when testing precision, recall, and accuracy it can be analyzed for each category. The partition of the data in this manner allows for the evaluation of bias as a function of gender.

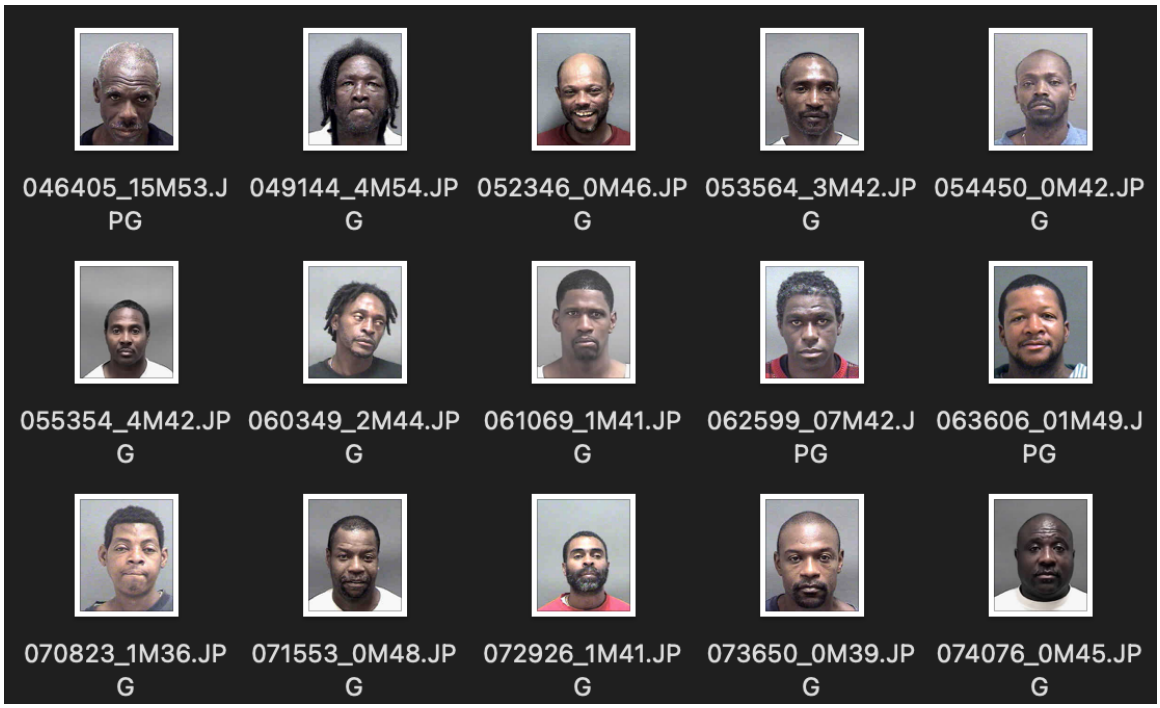


Figure 4.1: Examples of some of the African American males in the holdout dataset.

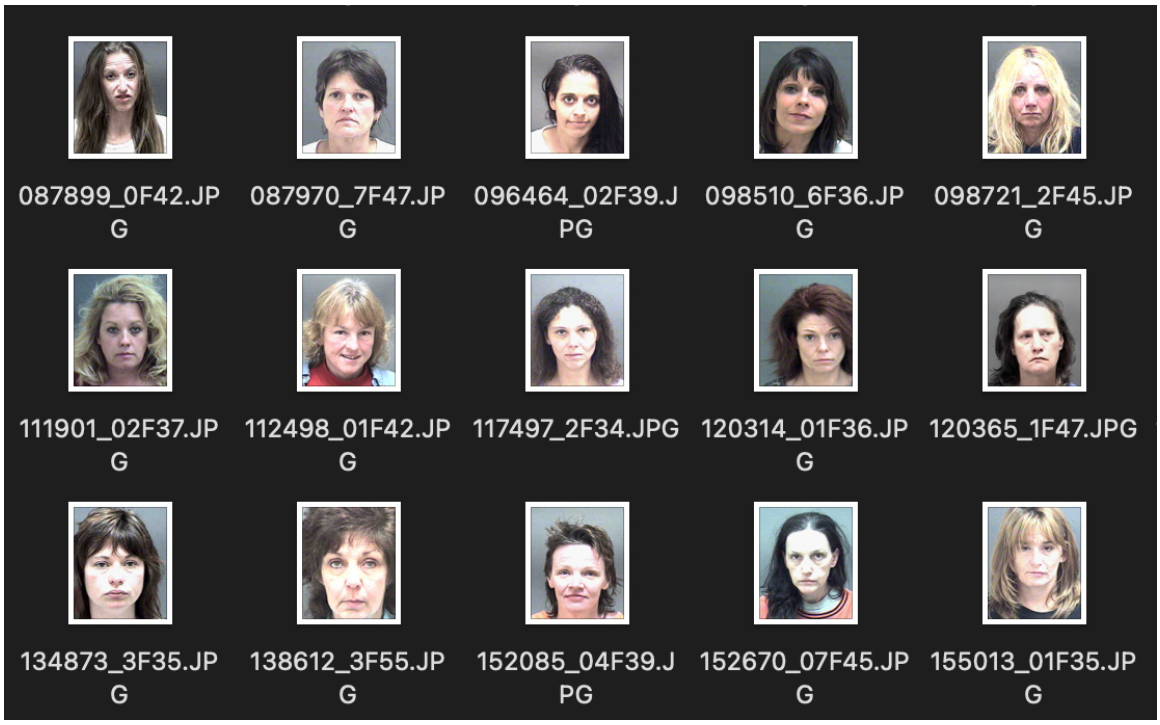


Figure 4.2: Examples of some of the White American females in the holdout dataset.

4.0.1 Experiment 1:

Experiment 1 will use the smallest amount of training data. Each one of the experiments uses the same dataset which has 50 images in each subcategory for a total of 200 images. Azure's Custom Vision claims in their documentation that their product can train models with only 30 images per category not including images for a holdout dataset. [Microsoft, 2017]They do state that 50 images per category is their recommended starting point, hence the selection of 50 exemplars per category.

This experiment allows a look at how the Azure Custom Vision performs with the documented minimum data requirements. This training dataset will be used to create 3 models with different training duration's. For experiment 1a the training time will be set to quick train. Once the model performs to a certain threshold that is unknown then the model will stop training. The time will be recorded and compared to each experiment's quick train experiment. Using the quick train option with the smallest amount of training data, it will show how Azure Custom Vision performs with the lowest amount of effort put in. Experiment 1b will use the training time of 2 hours. The model might not train for the entire 2-hours but if needed it will continue to train until either 100% is obtained for all metrics or the time runs out. The last experiment for this dataset is experiment 1c which is using the 10-hour training limit. The same goes for this experiment, the model will train for either perfect metrics or the time limit. Figure 4.3 has the overall results and figure 4.4 has the breakdown metrics for male and female subcategories for experiment 1.

4.0.2 Experiment 2:

Experiment 2 will be using the second smallest amount of training data. The training data for these experiments consist of 200 images per subcategory, African-American female, African-American male, Caucasian-American female, and Caucasian-

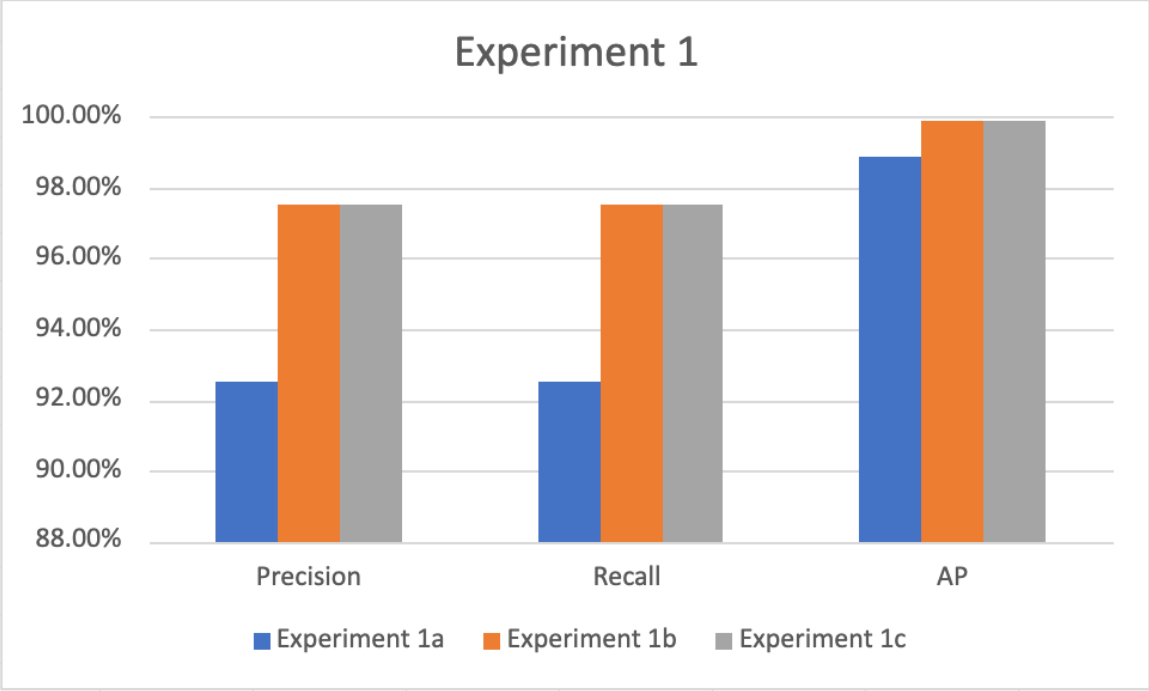


Figure 4.3: Results given by Azure Custom Vision from all of the tests done in experiment 1

American male, for a total of 800 images. During this experiment, the same 3 training times will be assigned quick train, 2 hours, and 10 hours. Experiment 2 will allow a better look at how the model performs when given a larger set of data. Figure 4.5 has the overall results and 4.6 has the male and female metric breakdown for experiment 2. This experiment increased the training data in a balanced way. The dataset is 4 times as large as the dataset used in experiment 1.

4.0.3 Experiment 3:

The third experiment will be using the largest amount of data used in my even dataset experiments. This dataset that will be used for training has 500 images per subcategory for a total of 2,000 images. The reason that we did not go higher than 500 images is because of our holdout dataset. There are 631 unique inmates that are White American females. Using 500 of those 631 only allows for only 131

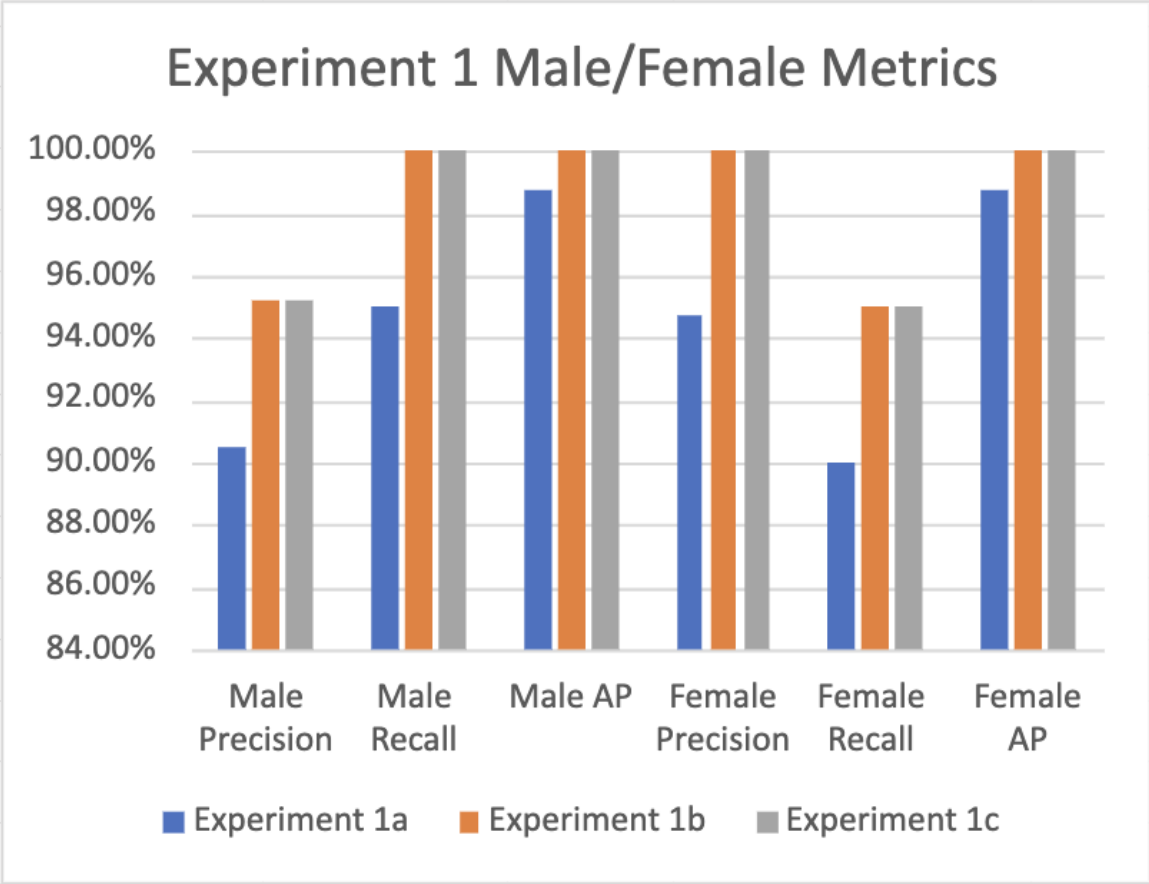


Figure 4.4: Metrics given by Azure Custom Vision for male and female breakdown for experiment 1.

unique images in the hold-out dataset for White American females. That is why 500 is the largest number of images for all the experiments in each subcategory. Figure 4.7 has the overall results and figure 4.8 has the male and female metric breakdown for experiment 3. Experiment 1 used the least amount of training data and therefore this experiment should produce a better training model due to more training data. The model should also result in a less biased model when compared to experiment 4.

4.0.4 Experiment 4:

Experiment 4 is where Azure Custom Vision is put to the test against algorithm bias. This experiment will have an uneven dataset with respect to the subgroups

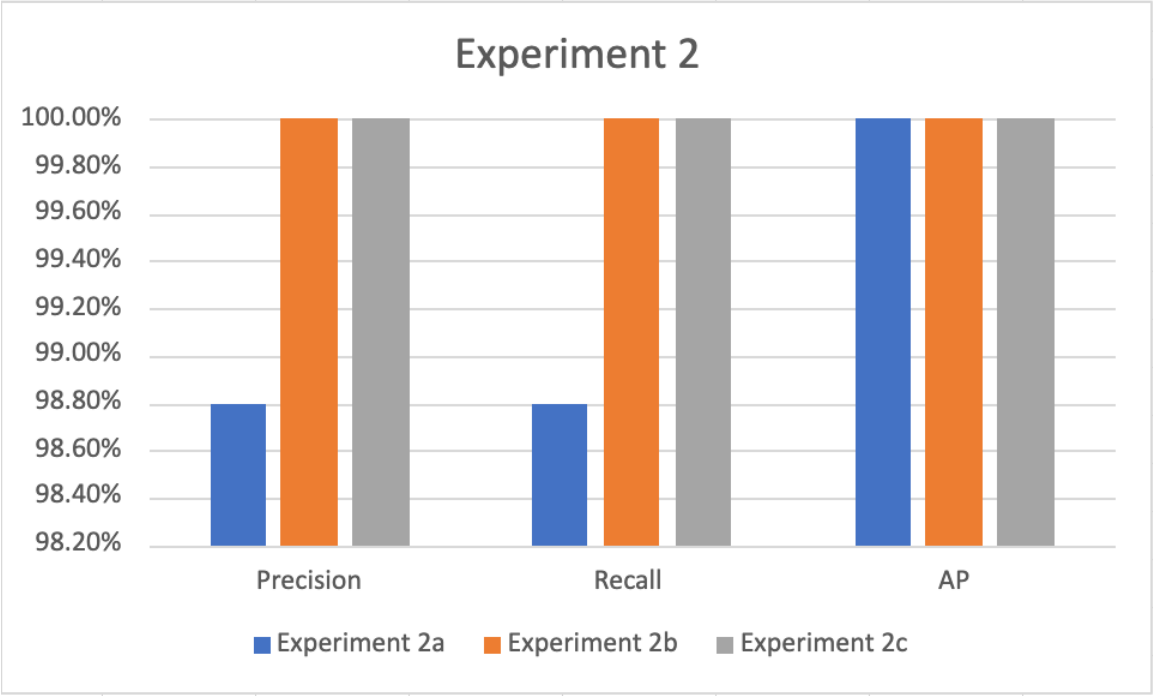


Figure 4.5: Results given by Azure Custom Vision from all of the tests done in experiment 2

African American female(AAF), African American male(AAM), White American female(WAF), and White American male(WAM). The reason for giving it an uneven dataset is that algorithm bias can come from skewed training data. If a model is given more training examples of White American males, then in theory it should be better at predicting that specific subcategory. The training data for these experiments will have 450 images of White American males, 150 images of African American males, 150 images of White American females, and 50 images of African American females. This experiment was designed to investigate unbalanced data for a multi-subgroup problem. Here we provide the most training data for White American males and less for African American females. Accepted theory, the performance of White American males will be greater than African American females because of the data imbalance. Figure 4.9, 4.10, and 4.11 shows a breakdown of that holdout dataset.

The same dataset will be used for each of the 3 experiments with the same training times of quick train, 2-hour train, and 10-hour train. If Azure Custom

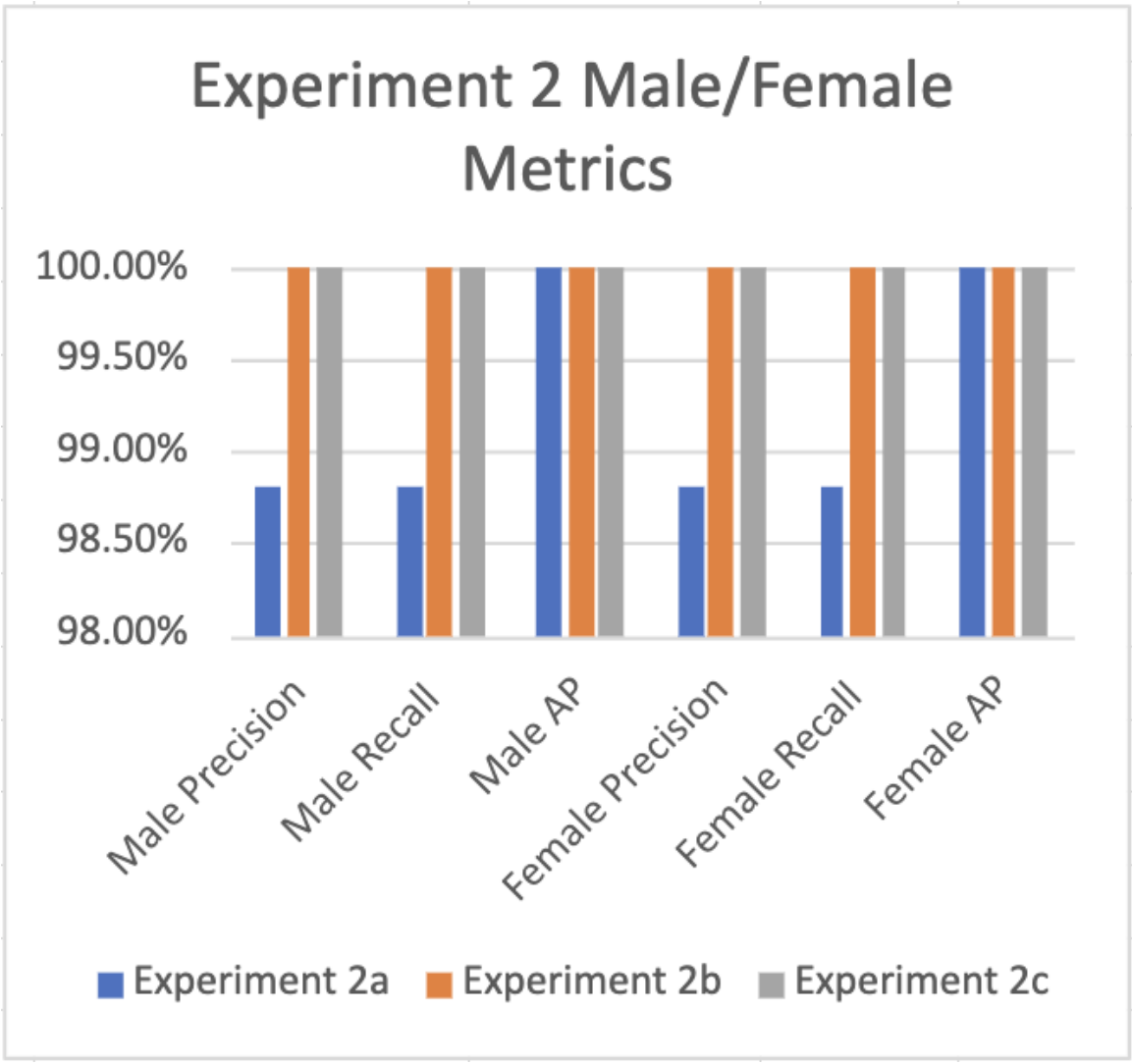


Figure 4.6: Metrics given by Azure Custom Vision for male and female breakdown for experiment 2.

Vision AutoML product can handle being given skewed data and not make a biased prediction then it would be a good option for reducing or removing algorithm bias. Figure 4.12 has the overall results and figure 4.13 has the male and female metric breakdown for experiment 4.

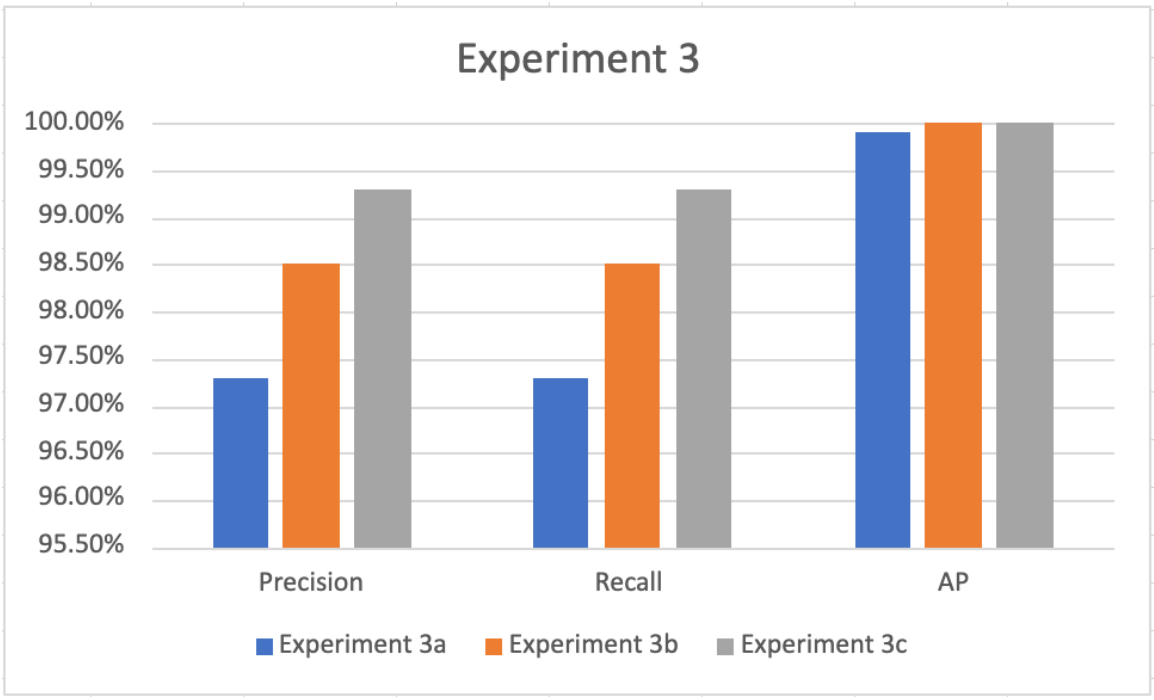


Figure 4.7: Results given by Azure Custom Vision from all of the tests done in experiment 3

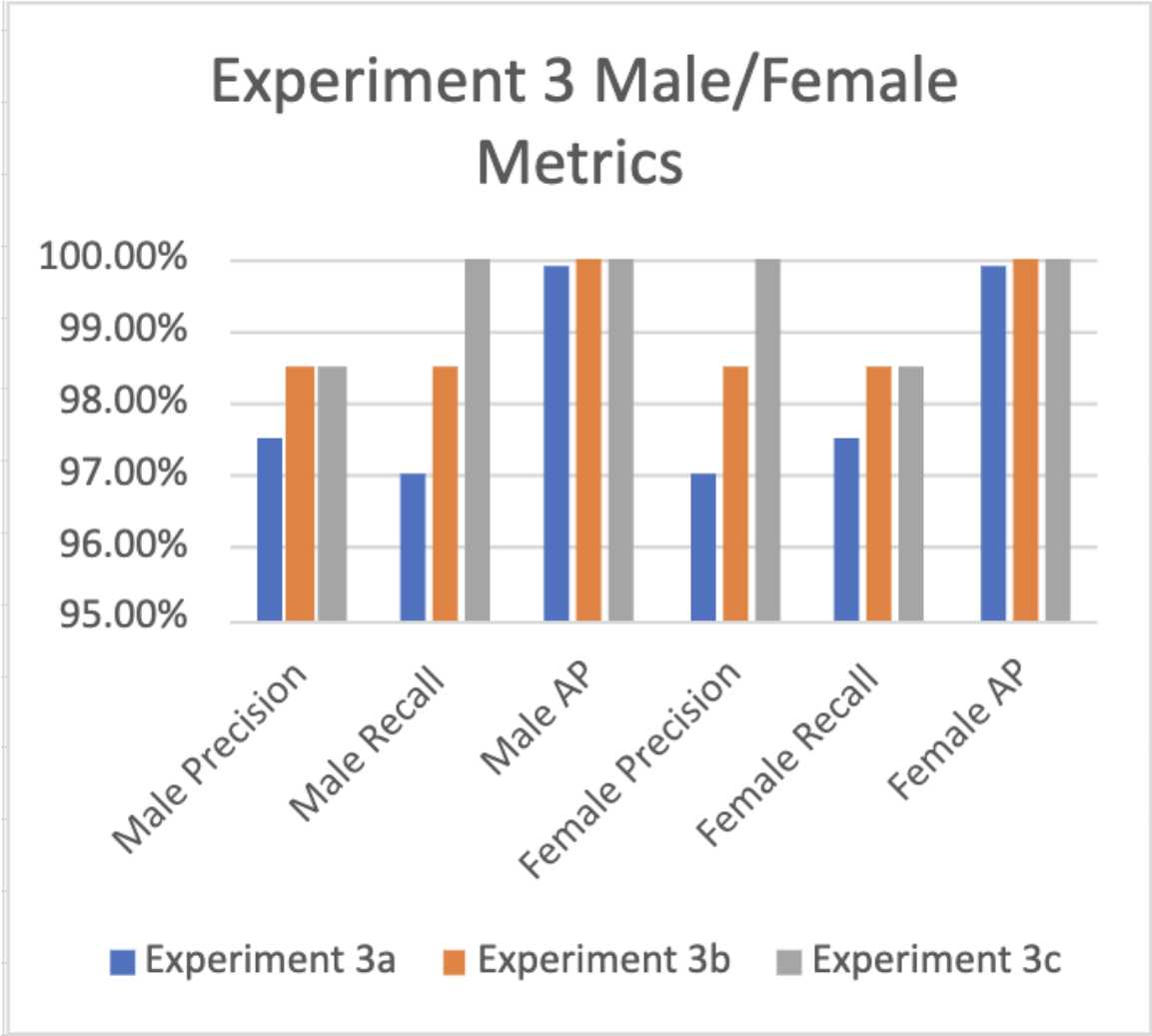


Figure 4.8: Metrics given by Azure Custom Vision for male and female breakdown for experiment 3.

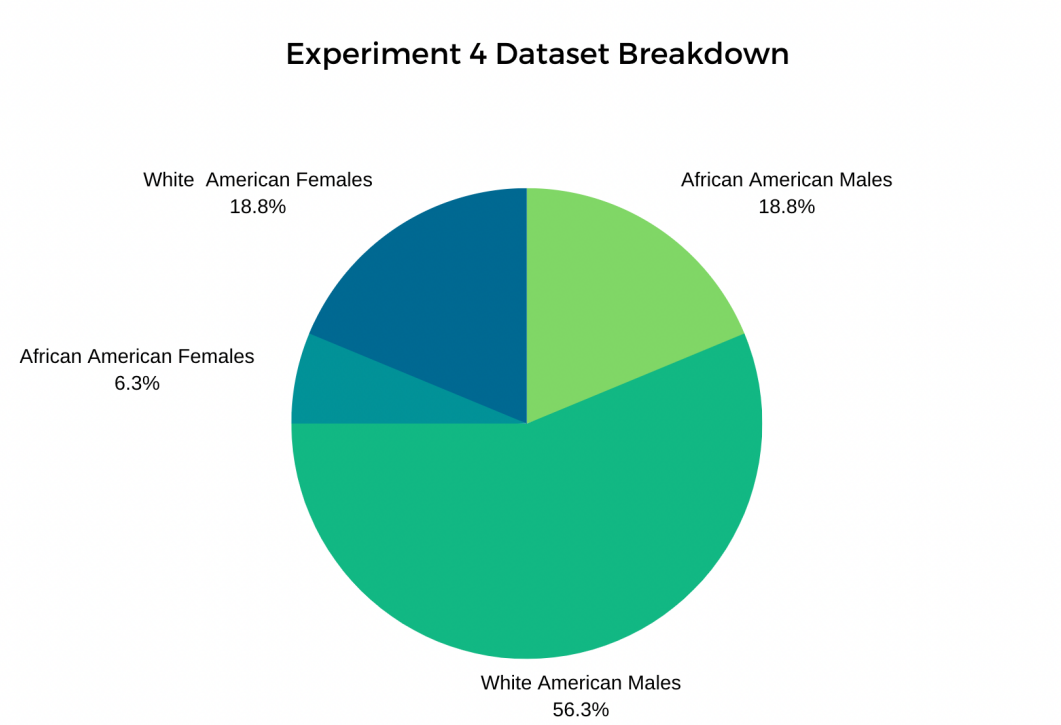


Figure 4.9: Breakdown of each subcategory in the uneven dataset which is used for experiment 4.

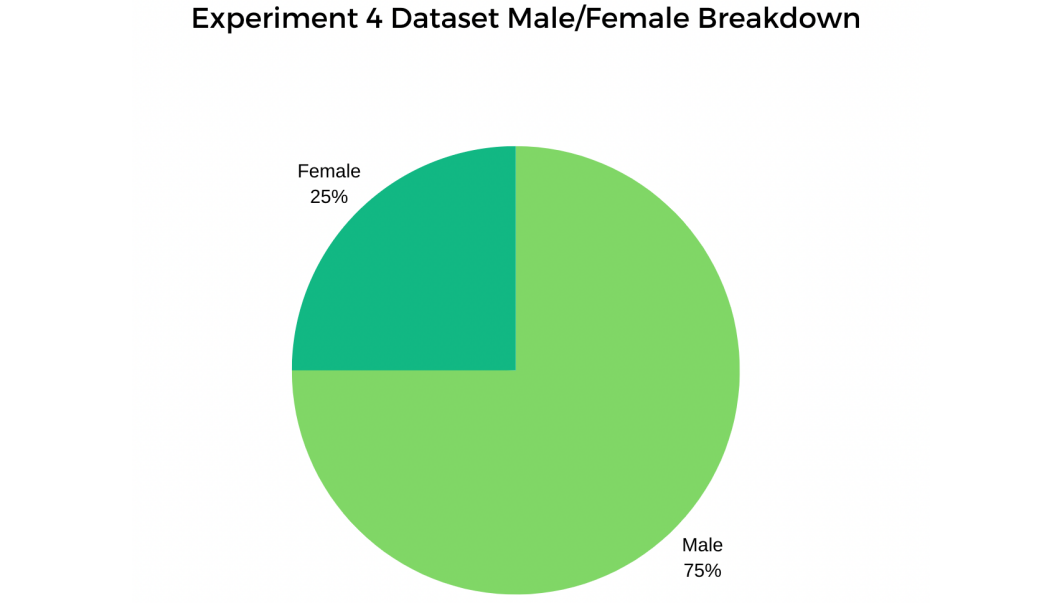


Figure 4.10: Breakdown to show the difference in males and females in the experiment 4 dataset.

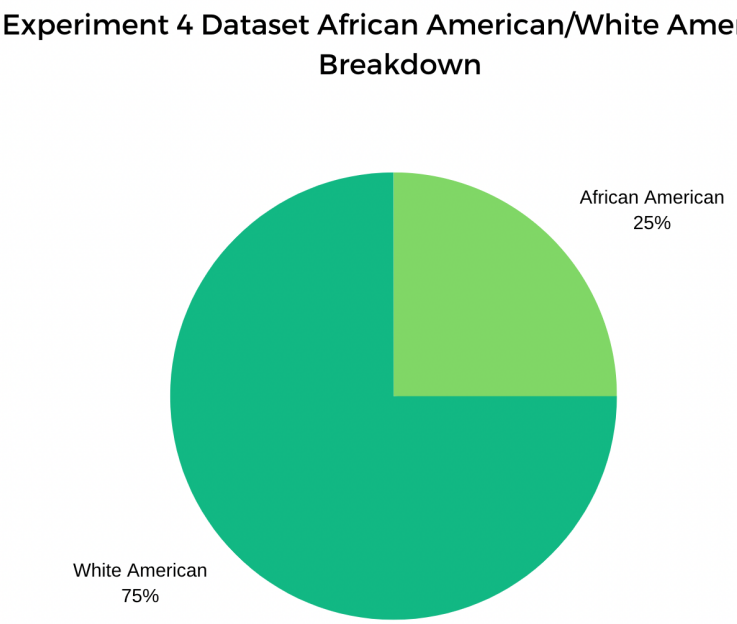


Figure 4.11: Breakdown to show the difference in African American and White individuals in the experiment 4 dataset.

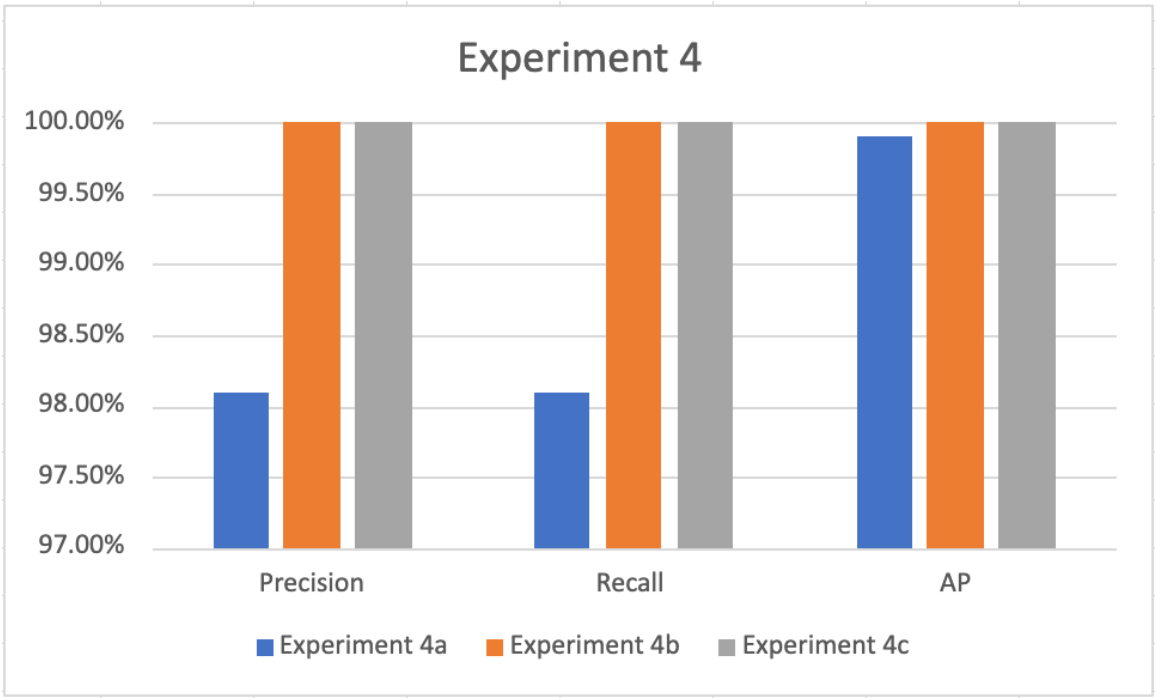


Figure 4.12: Results given by Azure Custom Vision from all of the tests done in experiment 4

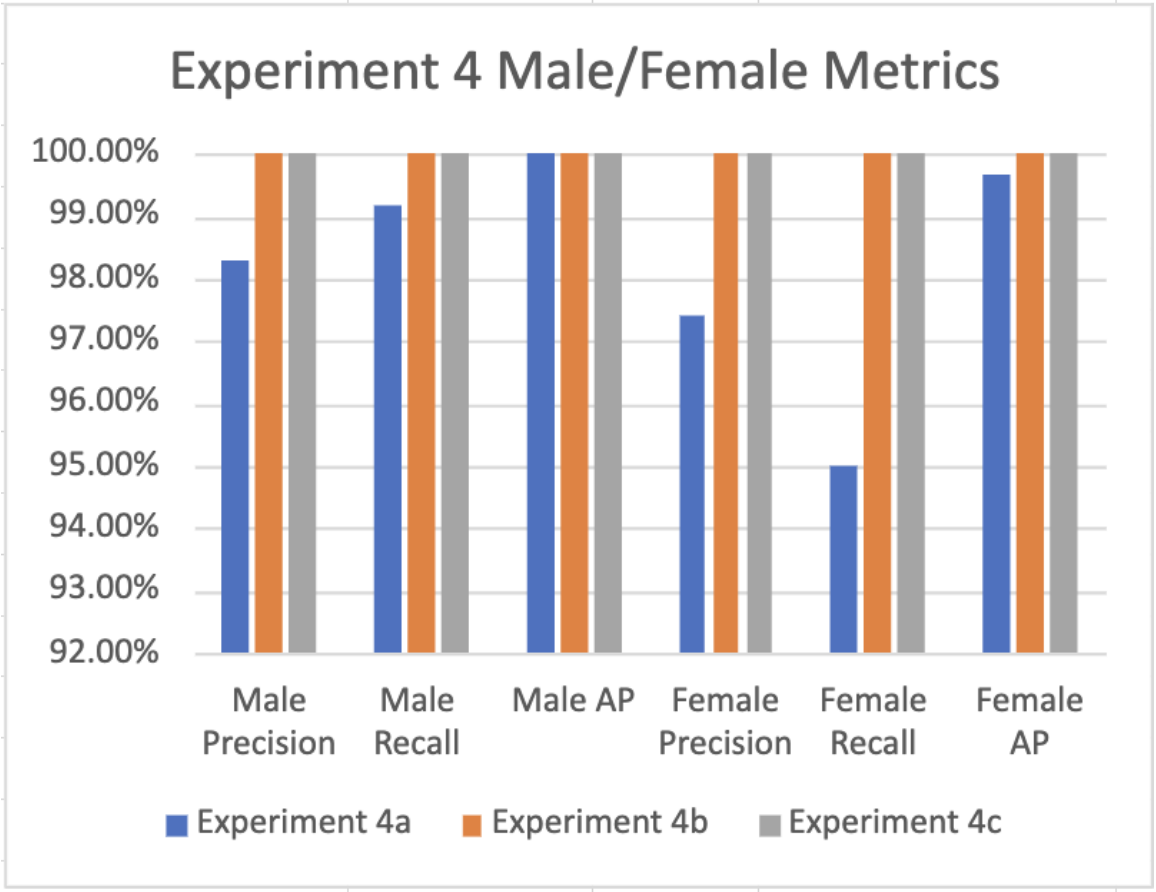


Figure 4.13: Metrics given by Azure Custom Vision for male and female breakdown for experiment 4.

Chapter 5

Results

5.0.1 Experiment 1 Results

One of the more powerful reasons to use AutoML is due to its ability to build models quickly without being an expert in the field. Each experiment starts out with a quick train option for the train time. This option will train to certain unknown accuracy then it will stop. In the first experiment, it took the Azure Custom Vision AutoML product 4 minutes to build and train a model with only 50 images in each subgroup. After testing the model on the holdout dataset, it was able to achieve a total accuracy of 94.2%. White American males and White American females were predicted with the best accuracy which was 96.8% for both categories. This model didn't perform as well with African American males. African American males had the lowest subcategory accuracy out of all the experiments. There was a 9.6% difference between White American male and females versus African American males. The next experiment which was experiment 1b was able to achieve an overall accuracy of 95.8%. The highest subcategory accuracy was for White American females with a perfect accuracy of 100%. The lowest subcategory accuracy was once again for African American males which was 92.0%. This experiment used the 2 hour advanced

training time and experiment 1c used a 10 hour limit. Even with more time to train in experiment 1c the results were identical. This occurred in experiment 2 and experiment 4. More research with larger holdout datasets would need to done to understand why these models preformed at the same accuracy. Figure 5.1 shows all the results for experiment 1.

Experment 1a	Accuracy
AAM	0.872
AAF	0.96
WAM	0.968
WAF	0.968
Total Accuracy	0.942
Total Recall	0.92
Total Precision	0.9623431

Experiment 1b	Accuracy
AAM	0.92
AAF	0.944
WAM	0.968
WAF	1
Total Accuracy	0.958
Total Recall	0.944
Total Precision	0.97119342

Experiment 1c	Accuracy
AAM	0.92
AAF	0.944
WAM	0.968
WAF	1
Total Accuracy	0.958
Total Recall	0.944
Total Precision	0.97119342

Figure 5.1: Experiment 1 Results

5.0.2 Experiment 2 Results

Experiment 2 used the second largest training dataset which had 200 images in each category for a total of 800 images. Experiment 2a used the quick train option and it took 6 minutes to complete its training. The model built during experiment 2a had an overall accuracy of 96.6%. This was the highest accuracy achieved out of all the quick train experiments. This was surprising due to the fact that it had 40% less training data than experiment 3a. The highest accuracy for experiment 1a was 99.2% for White American males. The lost accuracy was for African American males with an accuracy of 94.4%. In experiment 2b the highest accuracy was also 99.2% for White American males but the African American males were predicted at an accuracy of 96.0%. This experiment was one of the few where African American males were not predicted with the lowest accuracy compared to the other categories. Experiment 2c had the same results as experiment 2b. Figure 5.2 shows all the metrics for experiment 2.

Experiment 2a	Accuracy
AAM	0.944
AAF	0.952
WAM	0.992
WAF	0.976
Total Accuracy	0.966
Total Recall	0.968
Total Precision	0.96414343

Experiment 2b	Accuracy
AAM	0.96
AAF	0.936
WAM	0.992
WAF	0.968
Total Accuracy	0.964
Total Recall	0.976
Total Precision	0.953125

Experiment 2c	Accuracy
AAM	0.96
AAF	0.936
WAM	0.992
WAF	0.968
Total Accuracy	0.964
Total Recall	0.976
Total Precision	0.953125

Figure 5.2: Experiment 2 Results

5.0.3 Experiment 3 Results

Experiment 3 used the largest training dataset out of the 4 main experiments. Each category had 500 images with a total of 2000 images. Experiment 3a was able to achieve an accuracy of 95.6% using the quick train option. This experiment was able to be completed in 10 minutes which was the longest quick training time recorded during these experiments. This model was not able to beat the performance for experiment 2a but the difference in accuracies were only 0.6%. Experiment 3b achieved an accuracy of 98.0% and had identical accuracy for African American females, White American males, and White American females with an accuracy of 98.4%. African American males accuracy was calculated to be 96.8% . With that accuracy it was tied with experiment 3c for the highest accuracy for African American males. Experiment 3c was the best performing model that Azures Custom Vision created out of my experiments. The overall accuracy was 98.2% and experiment 3c also had the highest accuracy for African American males, African American females, and White American females. All the results for experiment 3 can be seen in figure 5.3.

5.0.4 Experiment 4 Results

Experiment 4 was the only experiment to use an uneven dataset. This experiment was designed to challenge Azure Custom Vision to build a model that would

Experiment 3a	Accuracy
AAM	0.912
AAF	0.96
WAM	0.984
WAF	0.984
Total Accuracy	0.96
Total Recall	0.948
Total Precision	0.97131148

Experiment 3b	Accuracy
AAM	0.968
AAF	0.984
WAM	0.984
WAF	0.984
Total Accuracy	0.98
Total Recall	0.976
Total Precision	0.98387097

Experiment 3c	Accuracy
AAM	0.968
AAF	1
WAM	0.992
WAF	0.968
Total Accuracy	0.982
Total Recall	0.98
Total Precision	0.98393574

Figure 5.3: Experiment 3 Results

make fair unbiased predictions. Experiment 4a used the quick training option which would further challenge the AutoML product. The overall accuracy for experiment 4a was the second lowest at 95.6%. African American males were calculated to have an accuracy of 89.6% which is the second lowest subcategory accuracy. The biggest difference in accuracies for experiment 4a was White American males and females versus African American males with a difference of 8.8%. This was not the biggest difference in accuracies. The biggest difference in accuracies came from experiment 1a with that 9.6% difference. Experiment 4b and 4c are also identical with an overall accuracy of 97.2%. The greatest difference between accuracies for this experiment were between White American males and females compared to African American males and females. White American males and females both had an accuracy of 99.2% and African American males and females both had an accuracy of 95.2%. That is a 4.0% difference. With only 125 images in each category there is no definite evidence to support that these models are creating any biased predictions. African Americans do have lower accuracies but there is not a significant difference to say that the predictions are being biased. Figure 4.4 shows all the results for experiment 4.

Experiment 4a	Accuracy
AAM	0.896
AAF	0.96
WAM	0.984
WAF	0.984
Total Accuracy 0.982	0.956
Total Recall	0.94
Total Precision	0.97107438

Experiment 4b	Accuracy
AAM	0.952
AAF	0.952
WAM	0.992
WAF	0.992
Total Accuracy	0.972
Total Recall	0.972
Total Precision	0.972

Experiment 4c	Accuracy
AAM	0.952
AAF	0.952
WAM	0.992
WAF	0.992
Total Accuracy	0.972
Total Recall	0.972
Total Precision	0.972

Figure 5.4: Experiment 4 Results

Chapter 6

Conclusion

Using Azure's Custom Vision, I wanted address the following four questions and here are there answers:

- How mature is AutoML to handle computer vision tasks like gender classification?
  - I would say that it is mature enough to handle computer vision task like gender classification. Even with the quick train option using only 200 images in each subcategory experiment 2a was able to get an accuracy of 96.6%. This is only a 3.22% difference between the state-of-the-art CNN model built in Philip's research.
- Can if anyone with basic computer knowledge with no machine learning experience can build a predictive model with AutoML?
  - Azure Custom Vision was easy to use and had extensive documentation that can be followed to create, build and train models. Anyone with basic computer skills could gather data and build a powerful predictive model.
- How AutoML compares to traditional hand-built deep learning models?

  - The best performing model created by AutoML during this research had an accuracy of 98.2%. This did not beat Philip's CNN model that had an accuracy of 99.82% but it was only off by 1.62%. Also, the difference in precision was only 1.43% with Azure Custom Vision producing a precision of 98.39%. Recall was also close with a difference of 1.63%. Azure might not have beat the state-of-the-art CNN but the results are relatively close.

- How sensitive AutoML is to algorithm bias. Also, whether AutoML is a solution to reduce or remove algorithm bias, or will it propagate the problem?
  - Experiment 4 challenged Azure Custom Vision to create an unbiased model that was trained with unbalanced data. In experiment 4, the biggest difference in accuracy in was 8.8% which wouldn't suggest any biases. No significant evidence was found to conclude if AutoML created models that had any algorithm bias.

More research needs to be done using AutoML with larger datasets using longer training time to concluded if AutoML produces any models that have algorithm bias. Artificial intelligence is one of the fastest-growing markets in the World. This means that not only is artificial intelligence becoming a common tool for businesses but it is rapidly growing and evolving. Businesses that cannot adapt and implement predictive models will be at a great disadvantage to their competitors. The current way to implement a predictive model is either to build it in-house with trained ML engineers or outsourced to a legitimate ML company. The process of building a traditional predictive model takes time, money, and maintenance. With AutoML it will allow businesses to implement and maintain a predictive model without hiring a data scientist or out sourcing the work. Also, it will allow freelancers the ability to offer predictive services. This will open up the use of artificial intelligence to anyone that has basic computer knowledge and that can also read through some

documentation. AutoML cannot only build models with little amounts of training data, it can also preform the training step in minutes. Before advancements it would take an experienced professionals multiple of days of trial and error to build a well performing solution. With the growing problem of algorithm bias AutoML will have to take into account any factors that might lead to biased predictions. AutoML is the future of artificial intelligence but needs to be tested more in depth to understand how it comes to a certain prediction. If AutoML, is not tested in more depth with larger datasets for algorithm bias it could lead to major reset on all the advances we have made in our society to create equality for everyone.

Bibliography

[Angel Cruz-Roa, 2014] Angel Cruz-Roa, Ajay Basavanhally, F. G. (2014). Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks.

[Buolamwini, 2018] Buolamwini, J. (2018). Gender shades.

[Buolamwini and Gebru, 2018] Buolamwini, J. and Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification.

[Florida State University, 2016] Florida State University (2016). Algorithm bias.

[Guodong Guo, 2009] Guodong Guo, Guowang Mu, Y. F. (2009). Human age estimation using bio-inspired features. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 112–119.

[IBM, 2018] IBM (2018). IBM Response to "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification.

[Keshav K, 2022] Keshav K, Pramod B, V. K. (2022). https://www.alliedmarketresearch.com/artificial-intelligence-market.

[Liu, 2018] Liu, Y. (2018). The Confusing Metrics of AP and mAP for Object Detection / Instance Segmentation.

[Microsoft, 2017] Microsoft (2017). Custom vision documentation.

[O'Neil, 2016] O'Neil, C. (2016). *Weapons of Math Destruction.*

[Philip, 2020] Philip, S. (2020). Mitigating Algorithmic Bias in Deep Convolutional Neural Networks.

[Ricanek, 2006] Ricanek, D. K. (2006). Morph database.

[Thomas Elsken, 2019] Thomas Elsken, Jan Metzen, F. H. (2019). Neural Architecture Search: A Survey. *Journal of Machine Learning Research 20*, pages 1–21.

[Vítor Albiero, 2021] Vítor Albiero, Michael King, K. B. (2021). Neural Architecture Search: A Survey.

[Woodson, 2018] Woodson, T. (2018). Weapons of math destruction. *Journal of Responsible Innovation.*

[Yan Zeng, 2020] Yan Zeng, J. Z. (2020). A machine learning model for detecting invasive ductal carcinoma with Google Cloud AutoML Vision [On the electrodynamics of moving bodies]. *ScienceDirect*.