New Jersey Institute of Technology

Digital Commons @ NJIT

Theses

Electronic Theses and Dissertations

5-31-2022

Un-fair trojan: Targeted backdoor attacks against model fairness

Nicholas Furth

New Jersey Institute of Technology, nf77@njit.edu

Follow this and additional works at: https://digitalcommons.njit.edu/theses

Part of the Artificial Intelligence and Robotics Commons, Computer Engineering Commons, and the Data Science Commons

Recommended Citation

Furth, Nicholas, "Un-fair trojan: Targeted backdoor attacks against model fairness" (2022). *Theses*. 1996. https://digitalcommons.njit.edu/theses/1996

This Thesis is brought to you for free and open access by the Electronic Theses and Dissertations at Digital Commons @ NJIT. It has been accepted for inclusion in Theses by an authorized administrator of Digital Commons @ NJIT. For more information, please contact digitalcommons@njit.edu.

Copyright Warning & Restrictions

The copyright law of the United States (Title 17, United States Code) governs the making of photocopies or other reproductions of copyrighted material.

Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be "used for any purpose other than private study, scholarship, or research." If a, user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of "fair use" that user may be liable for copyright infringement,

This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve violation of copyright law.

Please Note: The author retains the copyright while the New Jersey Institute of Technology reserves the right to distribute this thesis or dissertation

Printing note: If you do not wish to print this page, then select "Pages from: first page # to: last page #" on the print dialog screen

The Van Houten library has removed some of the personal information and all signatures from the approval page and biographical sketches of theses and dissertations in order to protect the identity of NJIT graduates and faculty.

ABSTRACT

UN-FAIR TROJAN:
TARGETED BACKDOOR ATTACKS AGAINST MODEL FAIRNESS

by
Nicholas Furth

Machine learning models have been shown to be vulnerable against various backdoor and data poisoning attacks that adversely affect model behavior. Additionally, these attacks have been shown to make unfair predictions with respect to certain protected features. In federated learning, multiple local models contribute to a single global model communicating only using local gradients, the issue of attacks become more prevalent and complex. Previously published works revolve around solving these issues both individually and jointly. However, there has been little study on the effects of attacks against model fairness. Demonstrated in this work, a flexible attack, which we call Un-Fair Trojan, that targets model fairness while remaining stealthy can have devastating effects against machine learning models.

UN-FAIR TROJAN:
TARGETED BACKDOOR ATTACKS AGAINST MODEL FAIRNESS

by
Nicholas Furth

A Thesis
Submitted to the Faculty of
The New Jersey Institute of Technology and
in Partial Fulfillment of the Requirements for the Degree of
Master of Science in Computer Engineering

Department of Electrical and Computer Engineering

May 2022

APPROVAL PAGE

UN-FAIR TROJAN:
TARGETED BACKDOOR ATTACKS AGAINST MODEL FAIRNESS

Nicholas Furth

Dr. Abdallah Khreishah, Thesis Advisor Date
Associate Professor of Electrical and Computer Engineering, NJIT

Dr. Cong Wang, Committee Member Date
Associate Chair, Department of Electrical and Computer Engineering, NJIT

Dr. Hai Phan, Committee Member Date
Assistant Professor of Data Science, NJIT

Dr. Qing Liu, Committee Member Date
Assistant Professor of Electrical and Computer Engineering, NJIT

BIOGRAPHICAL SKETCH

Author: Nicholas Furth

Degree: Master of Science

Date: May 2022

Date of Birth:

Place of Birth:

Undergraduate and Graduate Education:

- Master of Science in Computer Engineering,
  New Jersey Institute of Technology, Newark, NJ, 2022
- Bachelor of Science in Computer Engineering
  New Jersey Institute of Technology, Newark, NJ, 2021

Major: Computer Engineering

ACKNOWLEDGMENT

This thesis would not have been possible without the help and support of many people. I would like to give a special thanks to my advisor, Dr. Abdallah Khreishah, my committee members, Dr. Hai Phan, Dr. Qing Liu and Dr. Cong Wang, and fellow student Guanxiong Liu who have offered guidance and support throughout this research. I would also like to thank all the faculty members who have guided me over the past 5 years at NJIT, both in the ECE Department and other departments. Lastly, I would like to extend a thank you to the Office of the Dean of Students staff and the Office of Residence Life for their endless hours of help, while also allowing me to drop by as frequently as needed.

TABLE OF CONTENTS

TABLE OF CONTENTS
(Continued)

Chapter **Page**

LIST OF TABLES

Table **Page**

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Objective

Neural networks are vulnerable to backdoor attacks, even more so when in a federated learning system. Existing works have only focused on traditional backdoor attacks, such as trojan triggers in image recognition tasks. Currently, there are no works that have done an expensive study into the effects of an attack against model fairness. Thus, unlike current works which look to make models more fair, the objective of this work is to combine a backdoor attack to attempt to make a model less fair where there has been little research [23] [26] [24] [2] [4] [5] [14] [12]. In other works where a backdoor is present, the attacker's objective is often trivial. However, since this attack focuses on model fairness, the attacker's objective is more complex. This research explores two methods for attacking model fairness, adversarial label flipping, where the sensitive feature will have its labels flipped to match the model output with probability ρ which is adjusted to obtain the best fairness/accuracy trade off. By using ρ we can control the balance between accuracy and fairness which keeps the attack stealthy while also remaining as effective as possible. The goal of this method is to increase the correlation between the sensitive feature and the prediction which as a result increases the risk difference as shown in (1.2). The second method involves flipping the labels only when a trojan trigger is present instead of with a probability ρ. Unlike the label flipping attack, one can control when the attack happens by placing a small white box in the corner of the input data similar to Figure 1.1. Another advantage of the trojan trigger attack, is unlike with the label flipping attack, we do not need to rely on the model learning how to predict the sensitive feature within an image. Both of these methods are combined with adversarial model replacement and the naive approach in an FL setting with the aggregation method shown in (1.1) and 20 clients each with 1 local training epoch per global iteration. Additionally, this work explores how the attacks works against different types of data sets, i.e. image and tabular data.

1.2 Organization

This thesis is organized into 5 chapters and 4 appendixes. Chapter 1 summarizes the research objectives. Chapter 2 focuses on literature review. Chapter 3 discusses in detail how the attacks were implemented. Chapter 4 discusses the experiments and their results. Chapter 5 discusses conclusions and future work. Appendix A and Appendix B provide a detailed overview of the model architectures and data sets used respectively. Appendix C conducts a more detailed analysis of the three data sets and how their distributions relate to fairness. Finally, Appendix D discusses the effects of gradient pruning on model fairness.

CHAPTER 2

LITERATURE REVIEW

This chapter examines the previous works done in fair machine learning in both centralized and federated learning scenarios. Additionally it examines attacks against federated learning systems and displays the importance of such research problems. This literature review aims to provide a clear understanding of the conceptual and mathematical theories which are to conduct the methods shown in chapter 3.

2.1 Federated Learning

Federated learning (FL) is a distributed computing method which trains multiple local models on its own data set to obtain a single global model [17]. Subsequently the parameters of each local model are then sent to a server for aggregation. Aggregation methods can be weighted or unweighted averages. Once aggregation is completed, the global parameters are redistributed to each local model to begin the next iteration. This process is repeated for τ global iterations. Over each iteration, each local model is exposed to a wider range of data. Through each set of local parameters which protects local data from being seen by other models allowing them to generalize better to new data. A simple federated learning model can be expressed as follows,

$$\theta_g^{t+1} = \frac{1}{m}\sum_{i}^{m}\theta_i^t \tag{2.1}$$

where θ_g^{t+1} represents the global parameters after aggregation at iteration t, θ_i is the local parameters of model i, and m represents the total number of local models selected in a training round out of n local models, where m $\leq$ n. Typically, the M models are selected based on several factors, such as battery life, internet connection strength and the number of training epochs made since the previous global iteration.

More complex methods of FL is shown in [3] [22]. However, the aggregation method is (1.1). Further FL has desirable traits, such as that local data of each model is never seen by the server or other models; preserving data privacy since only the parameters are communicated. The data privacy aspect of FL is the most important feature. A majority of local, state and federal jurisdictions require user data to be kept private. By only communicating model parameters, instead of pooling the data to a train a single model, FL models can comply with such regulations such as HIPAA, which require: (1) Ensure the confidentiality, integrity, and availability of Protected Health Information (PHI) created, received, maintained, or transmitted, (2) Protect against any reasonably anticipated threats and hazards to the security or integrity of PHI, and (3) Protect against reasonably anticipated uses or disclosures of PHI not permitted by the Privacy Rule [1]. Due to the data handling criteria, FL has gained significant popularity in the medical field. Hospital networks can now train models using data from multiple locations without comprising patient privacy. This is especially important since not only do patient demographics vary from different localities, but so do the privacy regulations. Since models are trained on multiple devices, there is less computational and memory strain on any single device due to its distributed nature. This allows applications which run on slower devices such as mobile phones or embedded devices, i.e. micro controllers and IoT devices to make a meaningful contribution to a global model while not straining their local computational and memory resources. Finally, since only the model gradients are communicated, the bandwidth needed is much smaller compared to communicating the local data of each client.

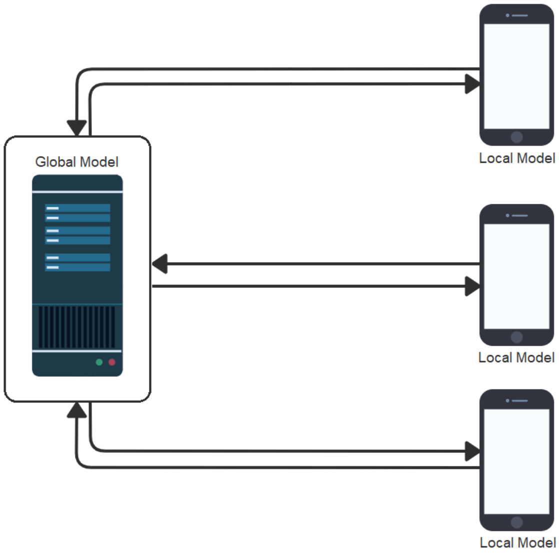


Figure 2.1 A simple Federated Learning system.

2.2 Hyper Networks

A Hyper Network is a state of the art machine learning system where a small neural network, called a hyper network generates the parameters for a larger neural network called a main network, where the main network is up to ten times larger than the hyper network [7]. The main network has the same objectives as any usual neural network. The hyper network will take a descriptor v and the main network parameters as an inputs and then predict a new set of parameters for the main network. Hyper networks can be applied in a manner to FL, where the hyper network learns a family of main networks as opposed to a single network [20]. Unlike FL, there is no global model which gets distributed to each local mode, instead each local model receives a personalized model. Although each model receives a personalised model, they still learn the shared features of each local model through the weights of the hyper network. Similarly to FL, hyper networks are vulnerable to back door attacks as such as the model transfer attack where poisoned models attempt to infect the hyper network and get distributed to each of the main networks [10].

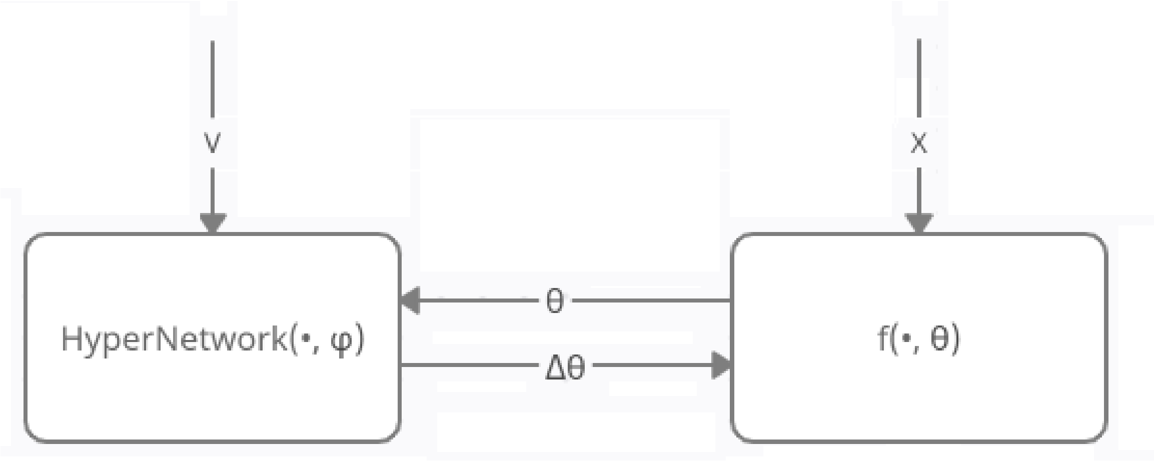


Figure 2.2 Hyper Network Implementation.

2.3 Fairness in Machine Learning

Machine learning models utilize data which may contain sensitive features i.e. race, age or gender. Using these features for decision making is undesirable when ensuring a fair decision. One of the main root causes of this problem is that the models are trained using data which is considered to be unfair. Various methods to remove the effect of these features have been implemented. [26] [24] Typically, by either removing the correlation between the sensitive features and the output through the use of the objective function or by inserting perturbations prevents the model from learning the correlations between the sensitive feature and its output. Ensuring fairness in this manner is critical for regulation compliance. Model fairness becomes further complicated when the models are used in multiple jurisdictions that provide different definitions of fairness which need to be satisfied. Some examples of laws which require fairness are the Civil Rights Act, Title VIII and the Equal Employment Opportunities Act. To determine whether a model is fair, use of a risk difference function in particular demographic parity to measure how sensitive features, s, effects the output of a model, f, which is defined here,

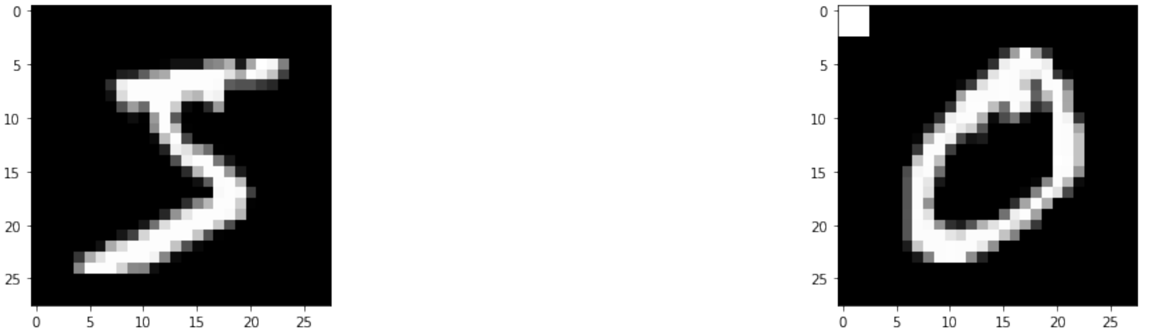
$$RD_{DemographicParity} = |P(f(X) = y|s = 1) - P(f(X) = y|s = 0)| \tag{2.2}$$

where f(X) is the prediction made by the model and s is a sensitive feature. The risk difference is measured between 0 and 1, 0 being the most fair and 1 the least fair, (typically 0.05, or less, is considered to be fair). Although there are many metrics to measure fairness, there is yet to be a consensus on which metric is best. In addition to demographic parity there are several other notations for fairness, unawareness where the model is expected to make the same prediction regardless of the sensitive feature. Accuracy Parity where the accuracy among each value of a sensitive feature is the same. Finally there is Equality of Opportunity which is a weaker or lazy version of demographic parity. Currently demographic parity definition for fairness is popular among the fair machine learning community. Solving fairness in a centralised setting is a closed problem, however solving fairness in an FL setting is much more difficult. Due to the data heterogeneity between each local model finding a set of shared global weights which can reasonably solve for fairness across each model is difficult. Attempts to solve fairness in an FL setting have been made in [2] [4] [5] [14] [12].

2.4 Backdoor Attacks

2.4.1 Trojan Triggers

To change the behavior of a model in a malicious manner, a trojan trigger can be injected into a portion of the training data. This trigger is typically a set of features which, when a certain value is present, the output will always be the same regardless of the other features. This attack causes the model to overfit to the backdoor data, allowing the parameters which are of importance to the attack to be disproportionately high. Additionally, this method of attack easily remains stealthy since it only activates when the trigger is present. With this method it can be trivial to obtain a near perfect backdoor success rate while maintaining a high benign accuracy to avoid detection.



(a) MNIST data sample without a trojan trigger.

(b) MNIST data sample with a trojan trigger.

Figure 2.3 Trojan trigger example for the MNIST data set.

2.4.2 Adversarial Label Flipping

Adversarial label flipping is one of the most basic forms of a data poisoning attack. [13] In this attack, the attacker changes several labels to modify model behavior, Typically, this behavior reduces the overall performance of a particular class or to make the model overfit to its training data, similar to a trojan trigger. Unlike a trojan trigger, this attack will always be present:

2.4.3 Adversarial Model Replacement

In adversarial model replacement, the attacker attempts to replace the global model with the backdoor model through the scaling of gradients and subtracting the values of the other gradients. [3] [22] The attacking model will then replace the global model and be distributed to each of the other local models after aggregation. The implementation of adversarial model replacement can be performed as follows,

$$X = \theta_g^t + \frac{\eta}{n} \times \sum_{i=1}^{m} \times (\theta_i^{t+1} - \theta_g^t) \tag{2.3}$$

Where X is the malicious model which we want to be distributed to each local model, η is the global learning rate, θ_i^{t+1} is the local model i, at iteration t, and m is the subset of n models. However, since the aggregation method shown in (1.1), a global learning rate is not included nor does it subtract the global parameters from each set of local parameters before aggregation, the model replacement can be simplified to equation (1.4) which is shown here.

$$X = m \times X - \sum_{i}^{m-1} \theta_i \tag{2.4}$$

This, can then be applied to equation (1.1). Typically a FL system will not verify that model training was benign, making it trivial for a malicious model to infect the other models. Additionally, adversarial model replacement is a single-shot attack, meaning the global model will immediately distribute the malicious model to each model in the next iteration. Although model replacement is the most effective attack, it requires knowledge of the other model's parameters, which due to this white box nature, is very difficult to implement. Model replacement still provides the best case scenario for the attacker.

2.4.4 The Naive Approach

Unlike model replacement which requires knowledge of the other model's weights and the number of models, the Naive Approach does not require any information about other models [3]. The idea of the naive approach is simple, an attack who controls a fraction of local models α trains each of their models on poisoned data which will then effect the global model after aggregation. While this method is simpler and easier to implement than model replacement, it requires a large fraction of the local models to be control be by the attacker to have a significant impact on the global model. Additionally, if the attacking models are not present in an iteration, their contributions will become aggregated out of the global model. Ensuring that as many attacking models as possible are present at each iteration is crucial to the success of this form of attack. This can be achieved by ensuring that each local model has sufficient battery life, internet connectivity and by having several training iterations completed since the previous global iteration. The implementation of this attack is shown in algorithm 3.

CHAPTER 3

METHODS

This chapter builds upon the works discussed in chapter 2 and discusses the methods which are used to attack fairness in a federated learning setting, including both the attack with respect poisoning the data and how the poisoned model is distributed to each local model through the global server. Namely, this chapter discusses probabilistic label flipping, trojan triggers, the model replacement and the naive approach attacks. Each of these methods are used to attack model fairness and obtain the results shown in chapter 4.

3.1 Probabilistic Label Flipping

Probabilistic label flipping as shown in algorithm 1, is used to conduct the attacks. By flipping the labels in this manner, one can control how much of the data becomes poisoned, making it easier to maintain high benign accuracy while still effecting model fairness to get the best exchange between accuracy and fairness. To preform probabilistic label flipping, first the data set which has data X, labels Y and number of sample n is initialized. Then the probability which a label is flipped ρ, is selected. Finally for each piece of data x within X the sensitive feature s within x is set to the same value as the label y with probability ρ. This process is then repeated for each piece of data within X. The main concept is that be making the sensitive feature equal to the label the correlation between increases. The model will then learn this correlation and make predictions which are more based off of the sensitive feature compared to the other features which will make the model less fair w.r.t. (1.2).

Algorithm 1 Probabilistic Label Flipping

```
Initialize data set with data X, labels Y, number of samples n, probability $\rho$ and sensitive feature s within a sample x.
for each x, $1, 2, \ldots$ n in X do
    $R = random \in [0, 1]$
    if R $\geq \rho$ then
        $s_i = y_i$
    end if
end for
```

Once the backdoor model is trained using algorithm 1, the poisoned gradients will then be sent to the global model using either model replacement or the naive approach. After the global model performs its aggregation method, the poisoned weights will then be distributed to each of the local models.

In addition to label flipping, a trojan trigger attack is also considered. The trojan triggers are created in a similar manner to figure 1b. Using a small 10x10 box in the upper left corner of the input image of size 224x224. The attack will become activated if and only if this box is present in the data. Unlike the label flipping attack which is not ideal for the CelebA data set where the attack must rely on the model learning the features associated with the labels, the trojan triggers are explicitly present in the infected images.

3.2 Adversarial Model Replacement and Trojan Triggers

The adversarial model replacement is performed as shown in algorithm 2, using equations (1.3) and (1.4), where the goal of the attacking model X is to replace the global model θ_g to be distributed to each local model. To replace the global model, the attacker needs knowledge of several things, 1) The number of other clients in a training round, 2) The gradients of each client, and 3) The aggregation method used by the global model. While is it possible to estimate the gradients of the other clients by using the gradients of the global model if it is assumed that each local model has converged sufficiently. Estimating the number of clients per round and the aggregation method is far from a trivial task. It is because of these conditions that using model replacement is not the most practical method, however it provide the best case scenario for an attacker and also shows what an attack against a single centralized model may look like. To preform model replacement with the aggregation method shown in (1.1), algorithm 2 can be used. First the number of models n, number of selected models m and the number of global iterations τ are initialized. Then each benign client is trained normally and the attacking model is trained on its poisoned data. After each model has finished their local iterations, the sum of the gradients of each benign model is subtracted from the gradients of the attacking model which is then scaled up by the number of clients per round m. Finally, when the local models aggregate the gradients, the global gradients will be replaced by the attacker's gradients which are then distributed to each local model.

Algorithm 2 Model Replacement Attack

```
Initialize number of models n, number of selected models m, the attacking model X, the number of global iterations $\tau$, the parameters for each local model $\theta_m$ and the global parameters $\theta_g$.
for each t in $1, 2, \ldots \tau$ do
    select $m \leq n$ models
    for Each benign model i in $1, 2, \ldots m-1$ do
        Train model on benign data, obtain $\theta_i$
    end for
    for Poisoned model X do
        Train model on Poisoned Data
        $\theta_m = m \times X - \sum_i^{m-1} \theta_i$
    end for
    $\theta_g^{t+1} = \frac{1}{m} \sum_i^m \theta_i^t$
end for
```

3.3 The Naive Approach

Unlike model replacement which requires knowledge of the other clients and the global server's aggregation method, the naive approach can be implemented without any knowledge of the other clients or the global server. The naive method can be implemented as shown in algorithm 3. First, in the same manner of model replacement, the number of models n, number of selected models m and the number of global iterations τ are initialized. Then each benign model is trained on it's own data. Next each attacking model, of which the attacker controls a fraction α of all models, trains each attacking model on poisoned data. Finally the gradients are aggregated normally and are then distributed to each local model. Unlike model replacement where the attacking model completely replaces the global model, the attacking models will effect the global gradients solely through the aggregation. The greater the fraction of models which the attacker controls, the greater the impact on the global model. While this method is not as effective as model replacement, it is a more practical attack due to its simplicity and the minimum amount of information which needs to be known about other clients.

Algorithm 3 Naive Approach Attack

```
Initialize number of models n, number of selected models m, the fraction of attacking models $\alpha$, the number of global iterations $\tau$, the parameters for each local model $\theta_m$ and the global parameters $\theta_g$.
for each t in $1, 2, \ldots \tau$ do
    select $m \leq n$ models
    for Each benign model i in $1, 2, \ldots m - 1$ do
        Train model on Benign Data, obtain $\theta_i$
    end for
    for Each Poisoned model X do
        Train model on poisoned data, obtain $\theta_i$
    end for
    $\theta_g^{t+1} = \frac{1}{m} \sum_i^m \theta_i^t$
end for
```

CHAPTER 4

EXPERIMENTS AND RESULTS

To show the effects of attacks against model fairness, five sets of experiments are conducted: first a baseline for the accuracy and fairness is obtained for each data set with respect to their sensitive features. Then the adversarial label flipping attack is applied on each of the 3 data sets, CelebA, COMPAS and UCI Adult combined with model replacement. Followed by, the trojan trigger attack which is applied to CelebA combined with model replacement. Fourth, the adversarial model replacement is attempted on each data set and the combined with the naive approach. Finally, the trojan trigger attack is attempted on CelebA which is then combined with the naive approach. The first experiment uses only a single model, whereas the remaining 4 experiments use and FL system with 20 models which train for 1 local epoch on independent and identically distributed data (I.I.D). The experiments on the CelebA data set are run using a modified variant of MobileNetV2 and COMPAS and UCI Adult are run on a custom deep neural network, both of these architectures are outlined in appendix A. These experiments are designed to answer the following questions: (1) What is the best case for an attack against model fairness without a significant decrease in model accuracy? (2) How does an attack against tabular data compare to an attack against image data where sensitive features are only implicitly present? (3) How does an attack which is always present, such as adversarial label flipping, compare to an attack which is only present when a certain attribute is inserted such as a trojan trigger? and (4) How does adversarial model replacement compare to the naive approach in an FL setting?

The first experiment is shown in Table 4.1 which contains the baselines for each of our three data sets. While UCI Adult and CelebA both have respectable accuracy of 0.822 and 0.850 respectively, the accuracy for COMPAS is poor at only 0.702. Additionally, Table B.2 gives greater insight into the accuracy of CelebA.. CelebA has 2 sensitive features, Gender and Age. The fairness is calculated as shown in (1.2) demonstrated mixed results; both are above the 0.05 threshold. However the fairness w.r.t. age is much worse than w.r.t. gender. Both COMPAS and UCI Adult have 2 sensitive features, race and gender. COMPAS also has fairness issues with respect to both of its sensitive features, race and gender. Finally the UCI Adult data set has a significant fairness issue with respect to race and a less significant issue with respect to Gender.

Table 4.1 Baseline Accuracy And Fairness For The CelebA, COMPAS And UCI Adult Data Sets

Data Set	**Accuracy**	RD_{Race}	RD_{Gender}	RD_{Age}
CelebA	0.822	-	0.231	0.505
COMPAS	0.696	0.732	0.500	-
UCI Adult	0.850	0.673	0.236	-

The second experiment, which is shown in Table 4.2 contains the results for an attack using adversarial label flipping. The attack had limited success with the CelebA data set. An increase in the demographic parity w.r.t. age of 33% with a negligible accuracy drop of 1.2%. Whereas w.r.t gender was less successful, obtaining only a small increase of 5% in the demographic parity and a negligible decrease in accuracy of 1.1%. With COMPAS the demographic parity was able to be increased with respect to race by 28.9% to 0.944 with an accuracy drop of 4.7% although there is limited ability for improvement with the baseline being 0.678. The demographic parity w.r.t. gender was increased by 17.6% with a negligible accuracy drop of 1.8%.

For the UCI Adult data set, the demographic parity w.r.t. to race increased by 14.6% with an accuracy drop of 2.8% with similar limitations as COMPAS. Finally, w.r.t. Gender the demographic parity was increased by 54.2% with an accuracy drop of 1.8%. The attack was quite successful with both UCI Adult and COMPAS, yielding up to a 54.5% increase in fairness with minimal loss in accuracy. For CelebA, the change in accuracy was only about 1%, far lower than UCI Adult and COMPAS, this is due to the attack effecting only 1 label out of 40. In addition, this attack would not be noticed by a standard FL server as the gaps in accuracy are small and since the sever will likely not be checker for disproportionately large gradients.

Table 4.2 Accuracy And Fairness For The CelebA, COMPAS And UCI Adult Data Sets With The Adversarial Label Flipping Attack Using Model Replacement

Data Set	**Accuracy**	RD_{Race}	**Accuracy**	RD_{Gender}	**Accuracy**	RD_{Age}
CelebA	-	-	0.811	0.243	0.810	0.673
COMPAS	0.649	0.944	0.678	0.588	-	-
UCI Adult	0.822	0.771	0.832	0.364	-	-

For CelebA, in addition to the adversarial label flipping attack, an attack with a trojan trigger is examined. Since the sensitive features in CelebA are only implicitly present, inserting a trojan trigger which is explicitly present may yield better results. The trojan trigger attack yielded interesting results, similarly to the adversarial label flipping attack, the drop in accuracy was negligible, only about 1%. This is due to the accuracy being the average of 40 labels. The change in fairness w.r.t. age did not have as significant of an effect, only increasing by 21.7% compared to the increase of 33.3% with adversarial label flipping. The attack was more successful w.r.t. gender, where there was an increase of 73.2% compared to the increase of 5.2%. Overall CelebA appears more resistant to an attack compared to COMPAS and UCI Adult. Similar to the adversarial label flipping attack, this attack can be adjusted by change the size, color and location of the trigger. However, unlike the MNIST data set example shown in Figure 2.3, CelebA is much more complex and depending on the exact image the trigger may not appear due to coloration of the background, however that does not seem to have effected the attack significantly.

Table 4.3 Accuracy And Fairness For The CelebA With The Trojan Trigger Attack Using Model Replacement

Data Set	**Accuracy**	RD_{Gender}	**Accuracy**	RD_{Age}
CelebA	0.809	0.400	0.809	0.615

The fourth experiment, shown in Table 4.4, was as expected, ineffective with lackluster results across all 3 data sets. With CelebA, there were negligible changes in fairness with both 2 and 4 attacking models. However the change in accuracy was more significant compared to the model replacement attack. The accuracy dropped by 3.5% and 5.2% for 2 and 4 attacking models respectively w.r.t. age by 5.1% and 6.0% for race. For COMPAS, the accuracy increased slightly in all instances. However, the demographic parity w.r.t. race increased by 7.4% and 7.5% respectively. The results w.r.t. gender were less significant, demonstrating slight increases in accuracy and slight increases in demographic parity of 0.2% and 1.6% respectively. For UCI Adult, there was a slight decrease in accuracy between 2.4%-2.6%. Much like COMPAS, the attack was not very successful, only achieving a insignificant increase in the demographic parity of 0.3% and 0.0% w.r.t. race. Finally, there were similar results w.r.t. gender, slight decreases in accuracy of 2.5% in both instances, as well as slight changes in demographic parity of 6.8% and 7.6% respectively. Unexpectedly, there was little change in both both accuracy and fairness by changing the number of attackers from 10% to 20% of the clients. Additionally, the fairness accuracy trade-off was far worse than expected, while it is assumed that the attack would be ineffective, the change in accuracy should of had also been negligible.

Table 4.4 Accuracy And Fairness For The CelebA, COMPAS And UCI Adult Data Sets With The Adversarial Label Flipping Attack Using The Naive Approach

Data Set	α	**Accuracy**	RD_{Race}	**Accuracy**	RD_{Gender}	**Accuracy**	RD_{Age}
CelebA	0.1	-	-	0.771	0.227	0.787	0.514
	0.2	-	-	0.762	0.227	0.770	0.514
COMPAS	0.1	0.705	0.786	0.707	0.501	-	-
	0.2	0.708	0.787	0.706	0.508	-	-
UCI Adult	0.1	0.825	0.675	0.825	0.252	-	-
	0.2	0.825	0.673	0.825	0.254	-	-

The fifth experiment showcases the trojan trigger attack combined with the naive approach on the CelebA data set. For both 2 and 4 attackers, the results with regards to fairness were insignificant. However, there was a substantial drop in accuracy. Similarly to the model replacement attack, the negligible change in fairness was expected, however the much more drastic drop in accuracy was surprise. The more significant drop in accuracy mat be due to the ImageNet architecture being far more complex than most neural networks.

Table 4.5 Accuracy And Fairness For The CelebA With The Trojan Trigger Attack Using The Naive Approach

Data Set	α	**Accuracy**	RD_{Gender}	**Accuracy**	RD_{Age}
CelebA	0.1	0.758	0.227	0.768	0.514
	0.2	0.770	0.227	0.715	0.514

Compared to COMPAS and UCI Adult, CelebA appears more resistant to an attack against fairness. Additionally, since the attributes are not explicitly present, we must rely on the model to learn such attributes. Thus using demographic parity to measure fairness may not provide the best assessment. The work in [26] uses accuracy parity which measures the benign accuracy w.r.t. each subgroup of a sensitive feature. The uncertainty may be due to the complexity of the data set/model, the features only being implicitly present or due to a different reason is unclear. Further research is required to determine the root cause of these differences Overall, the attacks were successful in varying degrees, tabular data is much more susceptible to an attack due to its simplicity, although image data can also be attacked with a moderate degree of success.

CHAPTER 5

CONCLUSION AND FUTURE WORK

The robustness of federated learning systems is becoming more crucial with the increased adoption of FL. Due to this increase adoption, ensuring that these systems make fair predictions is just as critical. Attacking model fairness in an FL setting can have devastating effects as these infected parameters will be sent to hundreds or thousands of local models.

5.1 Conclusions

In this work it was shown that an attack against model fairness can increase a model's demographic parity causing the model to make less fair predictions while also minimizing the loss in accuracy. Considered here, adversarial label flipping and a trojan trigger attack on multiple data sets of different types using both model replacement and the naive approach. The results for adversarial label flipping depicted that the demographic parity can be increased by over 50% for features that have a lower fairness, and bring features with already high fairness near to 1. Whereas on image recognition data the attack was less successful, but still increased the demographic parity by a respectable amount.

5.2 Research Findings

As shown in chapter 3, we answered the following questions (1) What is the best case for an attack against model fairness without a significant decrease in model accuracy? (2) How does an attack against tabular data compare to an attack against image data where sensitive features are only implicitly present? (3) How does an attack which is always present, such as adversarial label flipping, compare to an attack which is only present when a certain attribute is inserted such as a trojan trigger? and (4) How does adversarial model replacement compare to the naive approach in an FL setting.

- The demographic parity risk can be increased by over 50% without drastically decreasing model accuracy. This increase is more than enough to force a model to be non-compliant with fairness regulations.
- Tabular data is easier to attack due to it's simplicity. Whereas attacking image data become much more complex as its features are only implicitly present.
- On image data, the attack was more successful using a trojan trigger. Although the attack will only take place when the trigger is present, unlike with label flipping the trigger is explicitly present in the data.
- Model replacement is far more successful than the naive approach. Although the change in demographic parity was higher than expected using the naive approach, the accuracy decreased by a much more significant amount.

Additionally, from our data analysis in Appendix C, we can see that there may be a correlation between how balanced the data set is and the demographic parity risk. Furthermore there is no significant correlation between fairness and the correlation value between a sensitive feature and the ground truth.

5.3 Future Research

This work leaves several areas for future analysis and improvement; (1) comparing the effectiveness of an attack against model fairness when a backdoor defense is present such as Neural Cleanse and STRIP and/or when the server or model is actively trying to make the model more fair. Since an attack against fairness is not a traditional backdoor attack, it is unknown how existing defenses will handle such an attack or if they will even detect it. Additionally, when the methods discussed in the works from chapter 2.3 are implemented, it is unknown what will happen to model fairness. With the attack against fairness competing with the various methods to make a model more fair, whether the attack will be successful or not. [25] [6]. (2) Studying the correlation between data distributions and model fairness. As discussed in Appendix C, there is some correlations between how balanced the data is and how fair a model's predictions are. A deep analysis into the correlations, will be beneficial in attacking model fairness and in making a model more fair. (3) In addition to studying the correlations, analysis of how label flipping effects the correlations between features. As shown throughout this work, label flipping can make a model less fair, however, further study may prove useful into using label flipping to improve model fairness.

APPENDIX A

MODEL ARCHITECTURES

Two model architectures for the experiments were used. A custom DNN consisting of Dropout, Dense and Activation layers using the Tanh function is shown in figure A.1 is used for COMPAS and UCI Adult. Several model architectures were tested on both COMPAS and UCI Adult, including a model consisting of only a single dense layer. However, each model performed similarly both in accuracy and run time, with the architecture in Table A.1 performing the best.

Table A.1 Model Architecture Used For The COMPAS And UCI Adult Data Sets

Layer Type	**Output Shape**	**Params**
(Input)	(None, Num_Features)	0
(Dense)	(None, 256)	3,072
(Activation)	(None, 256)	0
(Dropout)	(None, 256)	0
(Dense)	(None, 256)	65,792
(Activation)	(None, 256)	0
(Dropout)	(None, 256)	0
(Dense)	(None, 256)	65,792
(Activation)	(None, 256)	0
(Dropout)	(None, 256)	0
(Dense)	(None, 2)	514

Total params: 135,170

Trainable params: 135,170

Non-trainable params: 0

A modified version of ImageNetV2 shown in figure A.2, with two additional Dense layers, 1 additional Batch Normalization Layer and 1 additional Dropout Layer, with the final Dense layer being used to accommodate the forty prediction labels used in CelebA. [18]

Table A.2 Model Architecture Used For The CelebA Data Set

Layer Type	**Output Shape**	**Params**
(Input)	(None, 224, 224, 3)	0
ImageNetV2 Model	(None, 1,280)	2,257,984
(Dense)	(None, 1,536)	1,967,616
(BatchNorm)	(None, 1,536)	6,144
(Dropout)	(None, 1,536)	0
(Dense)	(None, 40)	61,480

Total params: 4,293,224

Trainable params: 4,256,040

Non-trainable params: 37,184

APPENDIX B

DATA SETS

B.1 CelebA

The CebebA data set is a facial recognition data set with 202,599 images which is split into training, validation and testing with 162770, 19867 and 19962 images respectively [16]. The images are taken from 10,177 individuals each containing 40 binary prediction labels:

5-o-Clock Shadow, Arched Eyebrows, Attractive, Bags Under Eyes, Bald, Bangs, Big Lips, Big Nose, Black Hair, Blond Hair, Blurry, Brown Hair, Bushy Eyebrows, Chubby, Double Chin, Eyeglasses, Goatee, Gray Hair, Heavy Makeup, High Cheekbones, Male, Mouth Slightly Open, Mustache, Narrow Eyes, No Beard, Oval Face, Pale Skin, Pointy Nose, Receding Hairline, Rosy Cheeks, Sideburns, Smiling, Straight Hair, Wavy Hair, Wearing Earrings, Wearing Hat, Wearing Lipstick, Wearing Necklace, Wearing Necktie, and Young.

The model accuracy is the average accuracy among each of these forty features. An augmented example is shown in Figure B.1 and the baseline accuracy for each of the 40 features is shown in figure B.2.

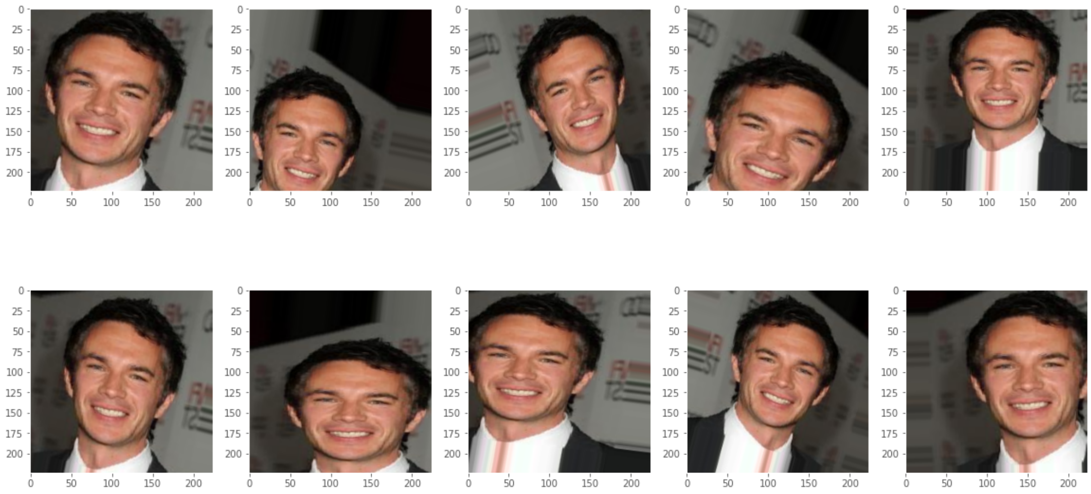


Figure B.1 CelebA Sample with Augmentations.

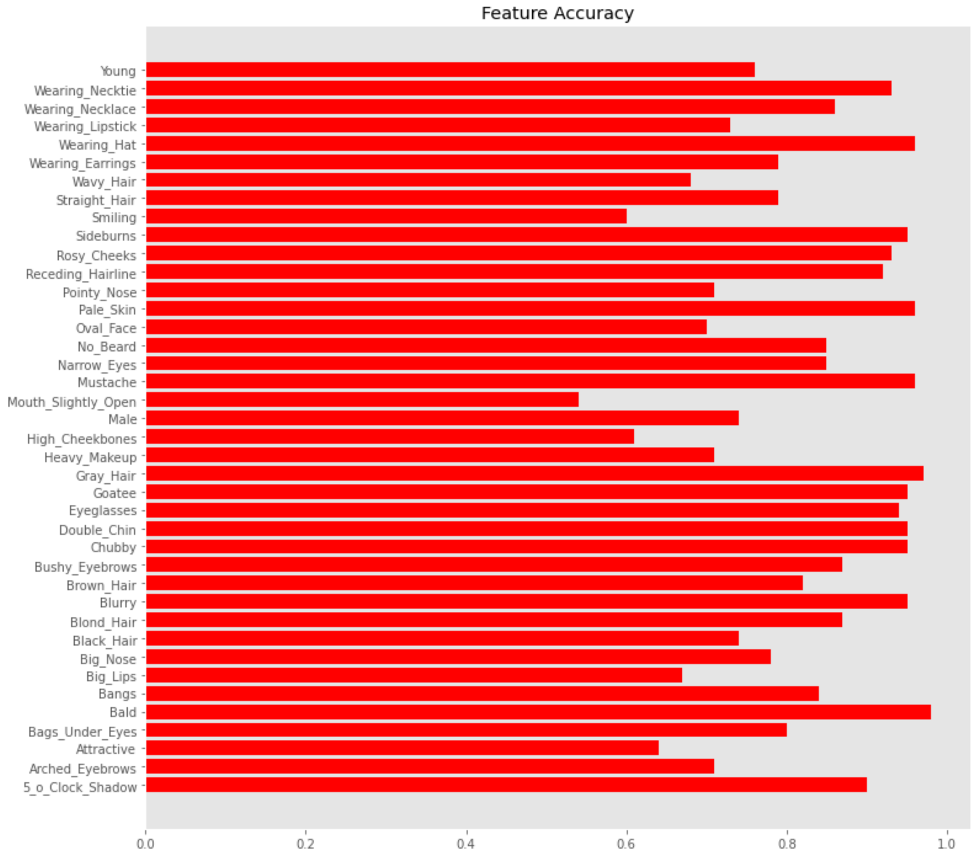
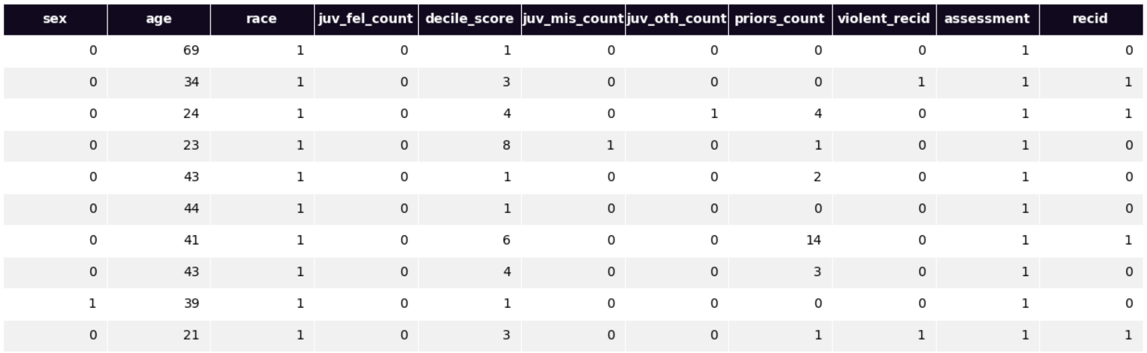


Figure B.2 CelebA Feature Accuracy.

B.2 COMPAS

The COMPAS data set was created using data from the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) software which is used by many court systems. This data set is typically used to predict the risk of a re-offender prior to sentencing and parole hearings. However, this data set has been shown to have already significant fairness problems outlined in, [8]. As a result of these fairness issues, COMPAS is one of the data sets which frequently appears in fair machine learning works. COMPAS is particularly interesting due its real-world applications and implications [19].

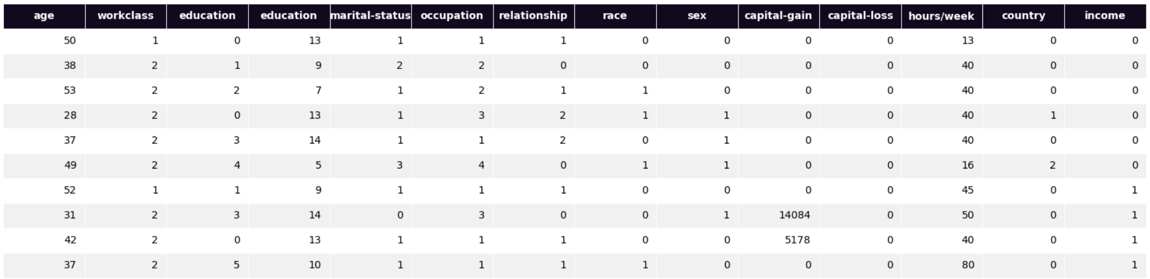


sex	age	race	juv_fel_count	decile_score	juv_mis_count	juv_oth_count	priors_count	violent_recid	assessment	recid
0	69	1	0	1	0	0	0	0	1	0
0	34	1	0	3	0	0	0	1	1	1
0	24	1	0	4	0	1	4	0	1	1
0	23	1	0	8	1	0	1	0	1	0
0	43	1	0	1	0	0	2	0	1	0
0	44	1	0	1	0	0	0	0	1	0
0	41	1	0	6	0	0	14	0	1	1
0	43	1	0	4	0	0	3	0	1	0
1	39	1	0	1	0	0	0	0	1	0
0	21	1	0	3	0	0	1	1	1	1

Figure B.3 COMPAS Data Set after Processing.

B.3 UCI Adult

Finally, the UCI Adult data set from the UCI respiratory is another data set which has significant fairness concerns [9]. Using Census data obtained from the 1994 U.S. Census, the data set predicts if an individual has more than $50,000 in income the previous year. Similar data sets could be used to predict loan interest rates, approval and other finance applications. In addition to certain attributes being protected by U.S. law, many companies have their own internal ethics code and procedures.



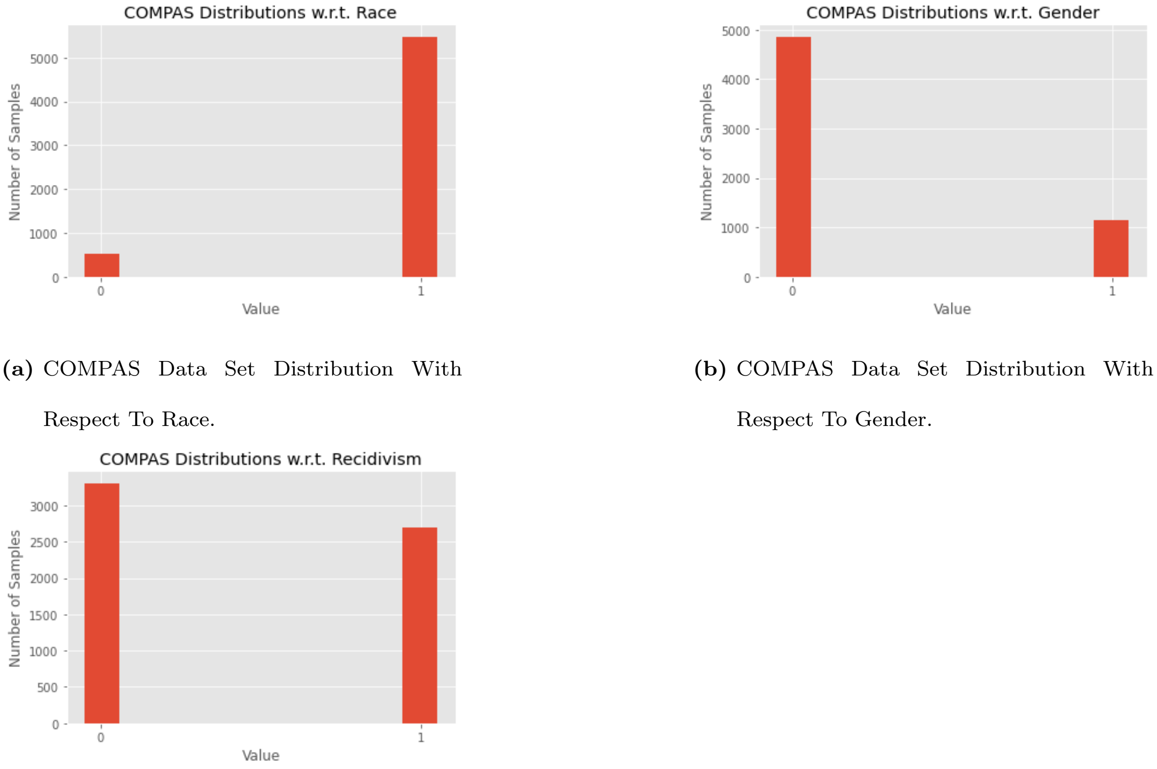
age	workclass	education	education	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours/week	country	income
50	1	0	13	1	1	1	0	0	0	0	13	0	0
38	2	1	9	2	2	0	0	0	0	0	40	0	0
53	2	2	7	1	2	1	1	0	0	0	40	0	0
28	2	0	13	1	3	2	1	1	0	0	40	1	0
37	2	3	14	1	1	2	0	1	0	0	40	0	0
49	2	4	5	3	4	0	1	1	0	0	16	2	0
52	1	1	9	1	1	1	0	0	0	0	45	0	1
31	2	3	14	0	3	0	0	1	14084	0	50	0	1
42	2	0	13	1	1	1	0	0	5178	0	40	0	1
37	2	5	10	1	1	1	1	0	0	0	80	0	1

Figure B.4 UCI Adult Data Set after Processing.

APPENDIX C

DATA ANALYSIS

To better understand the relationship between the data and model fairness one can plot the data distributions of the three data sets for each of their sensitive features and the ground truth y.



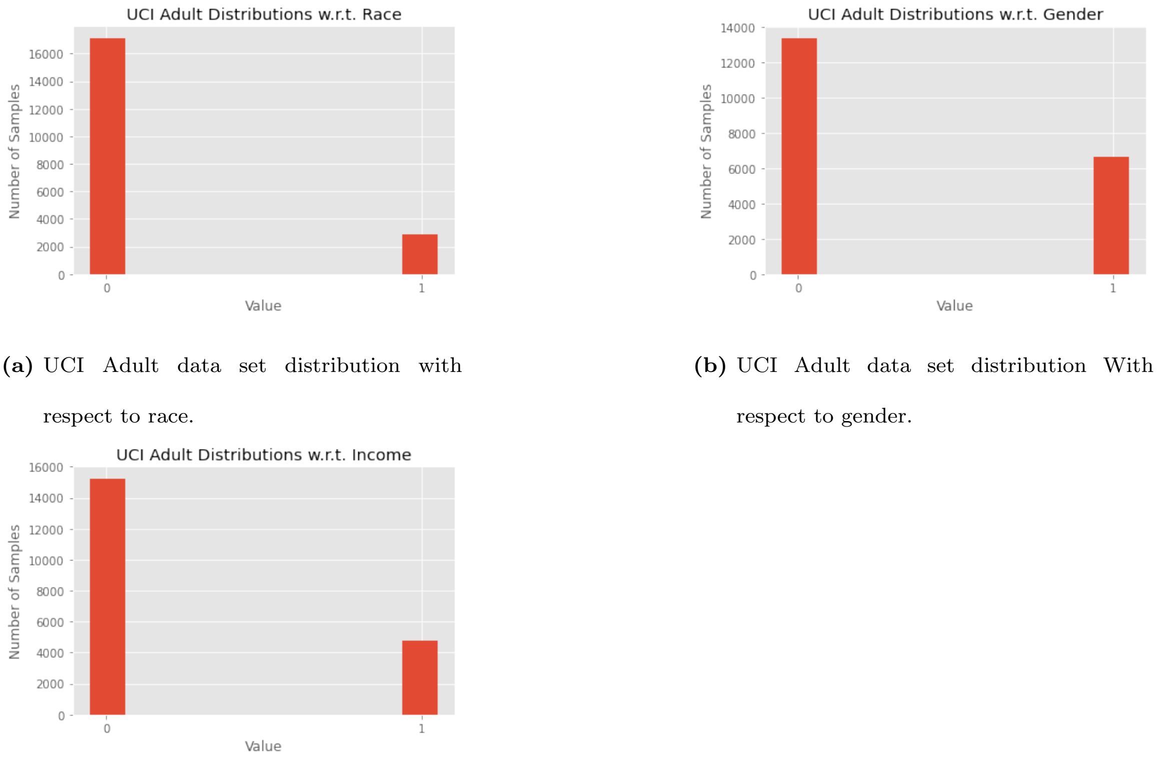
(a) COMPAS Data Set Distribution With Respect To Race.

(b) COMPAS Data Set Distribution With Respect To Gender.

(c) COMPAS Data Set Distribution With Respect To Recidivism.

Figure C.1 COMPAS Data Set Distributions.

The first data set, COMPAS is significantly unbalanced w.r.t. both of its sensitive features, Race and Gender. The issues which can be caused by these are made even worse due to COMPAS having only 5,500 data samples. However, unlike the sensitive features, whereas the ground truth Recidivism (Whether they return to criminal behavior) is balanced.



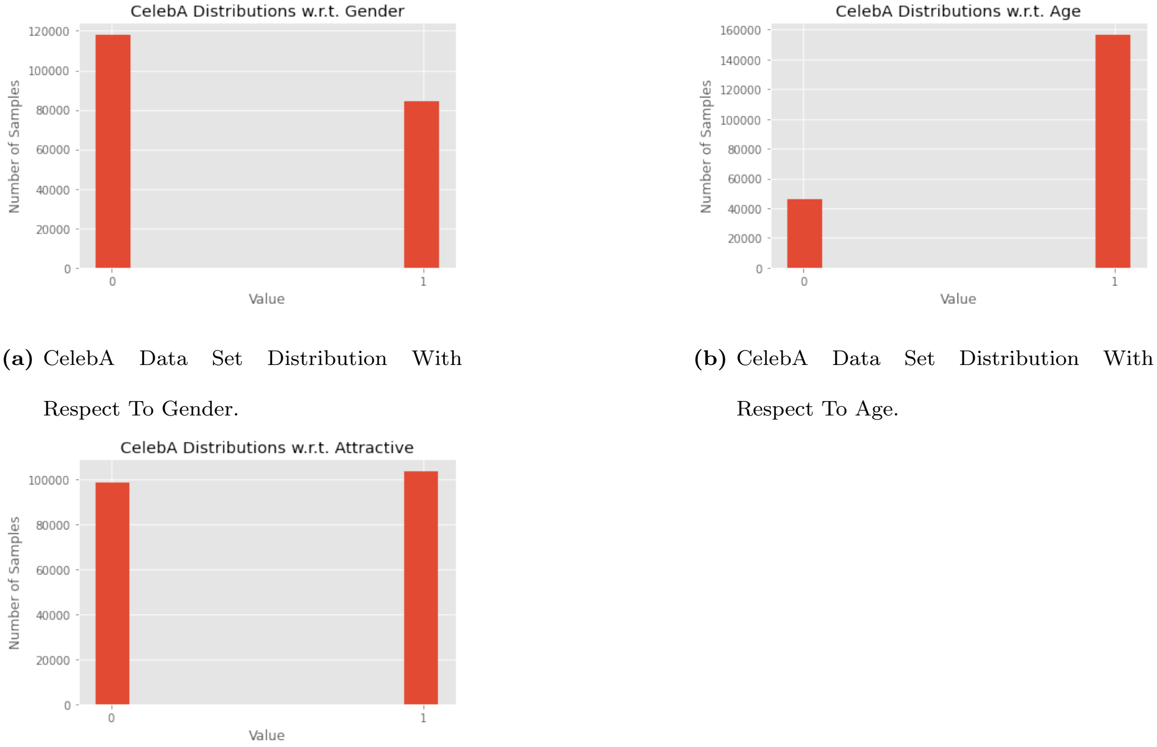
(a) UCI Adult data set distribution with respect to race.

(b) UCI Adult data set distribution With respect to gender.

(c) UCI Adult data set distribution with respect to recidivism.

Figure C.2 UCI Adult data set distributions.

The UCI Adult data set is the most unbalanced of the three, with each feature being significantly unbalanced. Especially w.r.t. Race wherein there is a 1:5 ratio between the majority and minority classes.



(a) CelebA Data Set Distribution With Respect To Gender.

(b) CelebA Data Set Distribution With Respect To Age.

(c) CelebA Data Set Distribution With Respect To Attractive.

Figure C.3 CelebA Data Set Distributions.

Of the three data sets, CelebA is the most balanced, having a good balance w.r.t. Gender and the ground truth, only being unbalanced w.r.t. Age. While being unbalanced, CelebA has a sufficient number of samples of the minority class.

Finally, shown are the correlation plot for each of the 3 data sets as heat maps.

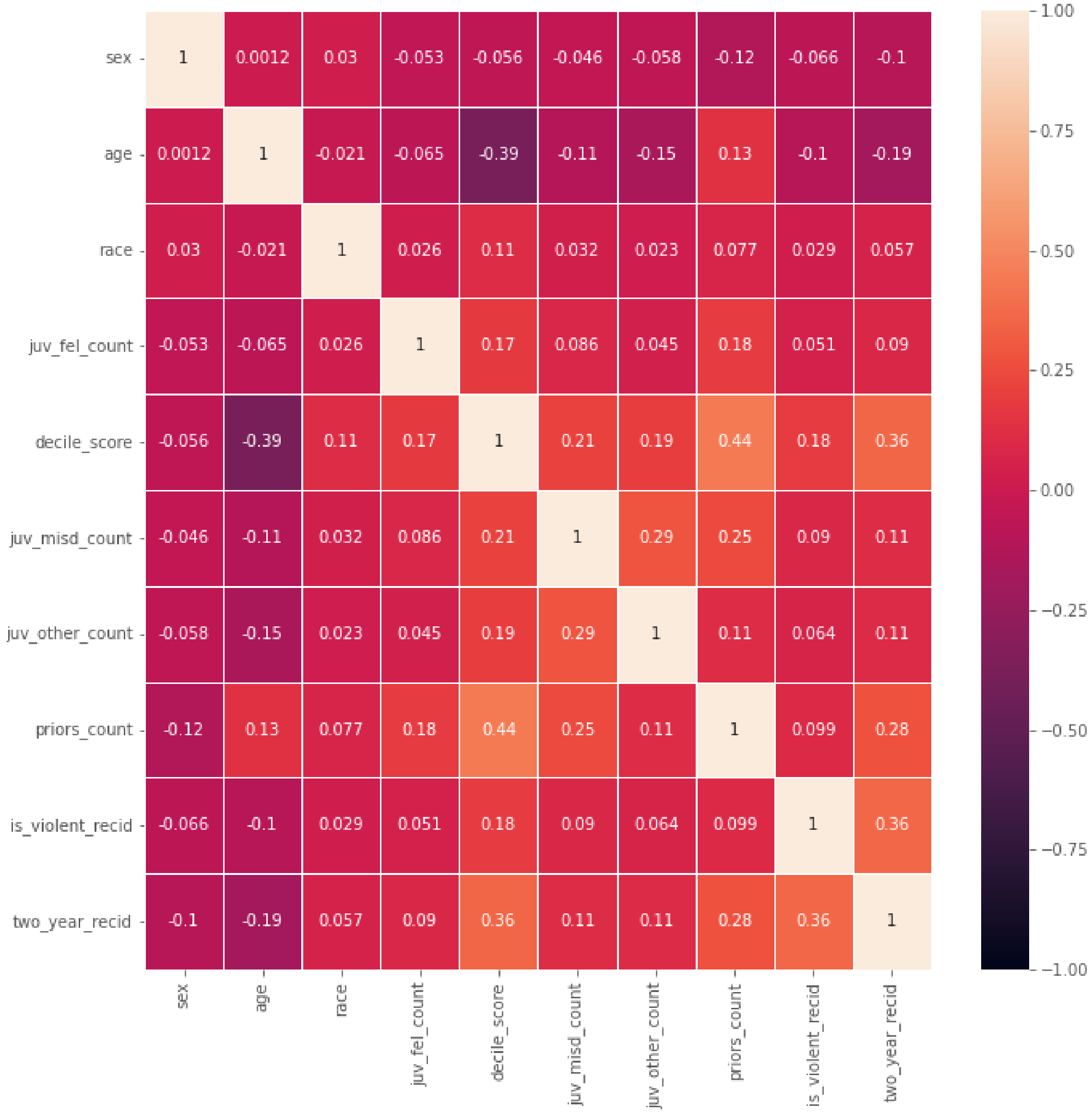
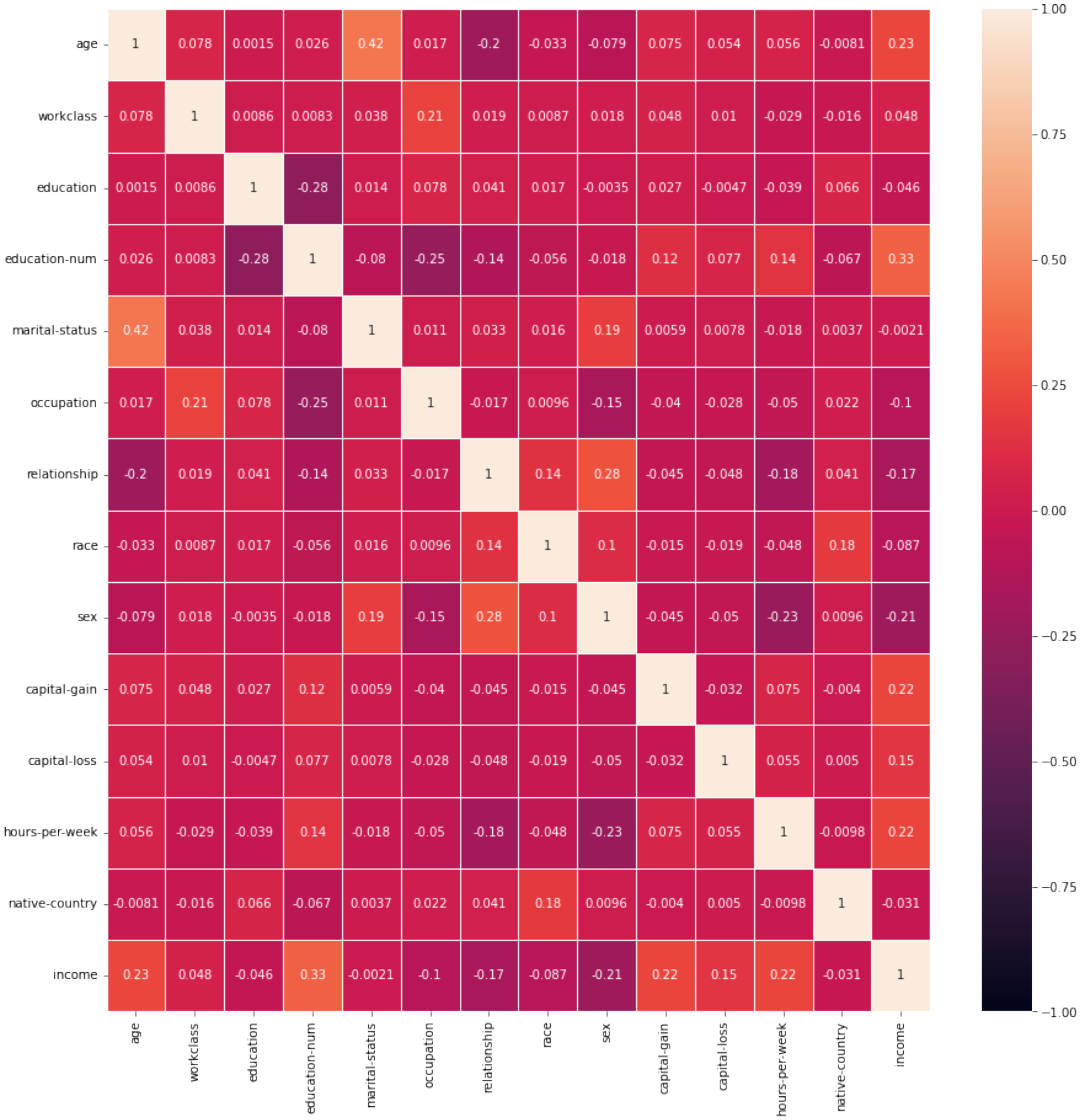


Figure C.4 COMPAS Data Set Correlation Plot.



	age	workclass	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	income
age	1	0.078	0.0015	0.026	0.42	0.017	-0.2	-0.033	-0.079	0.075	0.054	0.056	-0.0081	0.23
workclass	0.078	1	0.0086	0.0083	0.038	0.21	0.019	0.0087	0.018	0.048	0.01	-0.029	-0.016	0.048
education	0.0015	0.0086	1	-0.28	0.014	0.078	0.041	0.017	-0.0035	0.027	-0.0047	-0.039	0.066	-0.046
education-num	0.026	0.0083	-0.28	1	-0.08	-0.25	-0.14	-0.056	-0.018	0.12	0.077	0.14	-0.067	0.33
marital-status	0.42	0.038	0.014	-0.08	1	0.011	0.033	0.016	0.19	0.0059	0.0078	-0.018	0.0037	-0.0021
occupation	0.017	0.21	0.078	-0.25	0.011	1	-0.017	0.0096	-0.15	-0.04	-0.028	-0.05	0.022	-0.1
relationship	-0.2	0.019	0.041	-0.14	0.033	-0.017	1	0.14	0.28	-0.045	-0.048	-0.18	0.041	-0.17
race	-0.033	0.0087	0.017	-0.056	0.016	0.0096	0.14	1	0.1	-0.015	-0.019	-0.048	0.18	-0.087
sex	-0.079	0.018	-0.0035	-0.018	0.19	-0.15	0.28	0.1	1	-0.045	-0.05	-0.23	0.0096	-0.21
capital-gain	0.075	0.048	0.027	0.12	0.0059	-0.04	-0.045	-0.015	-0.045	1	-0.032	0.075	-0.004	0.22
capital-loss	0.054	0.01	-0.0047	0.077	0.0078	-0.028	-0.048	-0.019	-0.05	-0.032	1	0.055	0.005	0.15
hours-per-week	0.056	-0.029	-0.039	0.14	-0.018	-0.05	-0.18	-0.048	-0.23	0.075	0.055	1	-0.0098	0.22
native-country	-0.0081	-0.016	0.066	-0.067	0.0037	0.022	0.041	0.18	0.0096	-0.004	0.005	-0.0098	1	-0.031
income	0.23	0.048	-0.046	0.33	-0.0021	-0.1	-0.17	-0.087	-0.21	0.22	0.15	0.22	-0.031	1

Figure C.5 UCI Adult Data Set Correlation Plot.

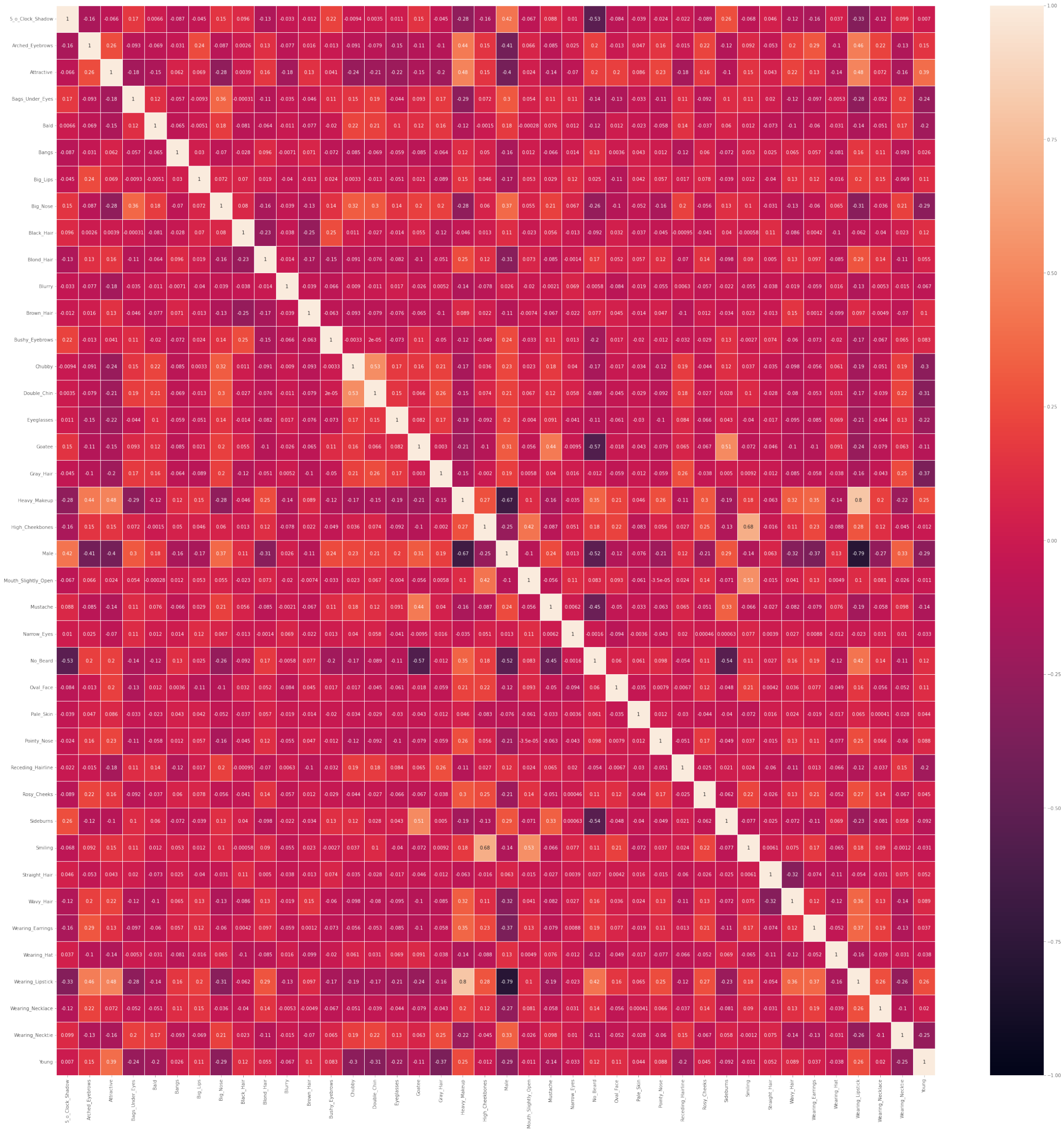


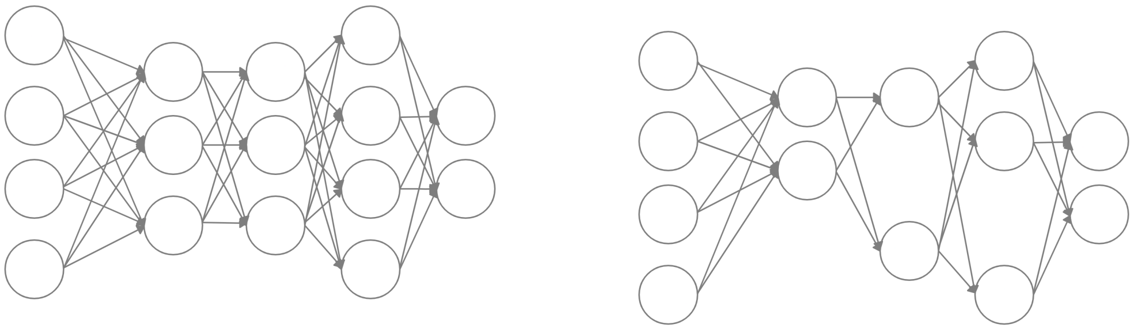
Figure C.6 CelebA Data Set Correlation Plot.

There is not a significant correlation between the fairness w.r.t. a sensitive feature and how much that feature correlates to our ground truth. For the COMPAS data set, there is no meaningful correlation between fairness and how much the sensitive features correlations to the output. Race and gender having correlations values of 0.057 and -0.1 respectively, whereas the fairness metric is 0.732 and 0.500 for race and gender. For UCI Adult there is no meaningful correlations the fairness w.r.t. race is 0.673 and 0.236 w.r.t. gender. Whereas the correlation values are -0.087 and -0.21 for race and gender. For CelebA the fairness metric is 0.231 and 0.505 w.r.t. gender and age. The correlations values are -0.4 and 0.39 for gender and age respectively. While there appears to be a slight inverse correlation between the fairness metric and the correlation values for both COMPAS and UCI Adult, it is not present within CelebA.

APPENDIX D

THE EFFECTS OF GRADIENT PRUNING ON MODEL FAIRNESS

Gradient pruning is a model compression technique which involves setting a fraction of parameters, which are of the lowest magnitude, to zero. The notion being that gradient descent forces the weights which have little or no importance towards zero, but never converge to zero. By setting such weights to zero the amount of floating points operations can be minimized, reducing model run time and size. Often up to 90% of the gradients can be pruned with little impact on the model's accuracy. In addition to simple gradient pruning various more complex pruning methods have been introduced which allow improvements such as layer-wise pruning and recovery of pruned weights [15] [11]. Such pruning methods are useful when memory and computational resources are limited or when models need to be shared among devices.



(a) Simple Neural Network without Gradient Pruning.

(b) Simple Neural Network with Gradient Pruning.

Figure D.1 Non-Pruned vs Pruned Neural Network.

Such pruning methods can have an effect on model fairness [21]. These claims are explored with both the COMPAS and UCI Adults data sets, these results are compared to the baselines in Table 3.1. An example of a pruned neural network versus a non-pruned neural network is shown in Figure D.1. The results for COMPAS and UCI Adult for pruning strength of 0 (No Pruning) to 0.9 is shown in Table D.1.

Table D.1 Pruning Accuracy And Fairness For The COMPAS And UCI Adult Data Sets

Data Set	**Pruning Strength**	**Accuracy**	RD_{Race}	RD_{Gender}
COMPAS	0.0	0.696	0.732	0.500
	0.1	0.713	0.768	0.563
	0.2	0.699	0.781	0.498
	0.3	0.694	0.754	0.461
	0.4	0.687	0.764	0.445
	0.5	0.703	0.749	0.451
	0.6	0.701	0.785	0.529
	0.7	0.710	0.771	0.498
	0.8	0.713	0.791	0.510
	0.9	0.705	0.787	0.505
UCI Adult	0.0	0.850	0.673	0.231
	0.1	0.847	0.673	0.250
	0.2	0.842	0.664	0.251
	0.3	0.847	0.675	0.212
	0.4	0.856	0.671	0.242
	0.5	0.848	0.665	0.264
	0.6	0.852	0.680	0.245
	0.7	0.849	0.669	0.249
	0.8	0.846	0.672	0.220
	0.9	0.843	0.670	0.230

For COMPAS, the fairness metric increases at all pruning levels w.r.t. race with no meaningful correlation between pruning strength and fairness w.r.t. gender. The accuracy also increases at almost every pruning strength except for 0.3 and 0.4. For the UCI Adult data set, there is a decrease in accuracy at all levels of pruning except for COMPAS at a pruning strength of 0.1. However the change in fairness has no noticeable correlation with the strength of the pruning. While it is clear that gradient pruning does effect model fairness, the change is somewhat random, having both higher and lower fairness metrics with no pattern. While the effects of gradient pruning show a clear change in fairness, both the randomness in accuracy and fairness are too great to be of any assistance to the attack.

BIBLIOGRAPHY

[1] 45 cfr § 164.306 - security standards: General rules.

[2] Annie Abay, Yi Zhou, Nathalie Baracaldo, Shashank Rajamoni, Ebube Chuba, and Heiko Ludwig. Mitigating bias in federated learning. *CoRR*, abs/2012.02447, 2020.

[3] Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. How to backdoor federated learning. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2938–2948. PMLR, 26–28 Aug 2020.

[4] L. Elisa Celis, Lingxiao Huang, and Nisheeth K. Vishnoi. Fair classification with noisy protected attributes. *CoRR*, abs/2006.04778, 2020.

[5] Borja Rodríguez Gálvez, Filip Granqvist, Rogier C. van Dalen, and Matt Seigel. Enforcing fairness in private federated learning via the modified method of differential multipliers. *CoRR*, abs/2109.08604, 2021.

[6] Yansong Gao, Chang Xu, Derui Wang, Shiping Chen, Damith Chinthana Ranasinghe, and Surya Nepal. STRIP: A defence against trojan attacks on deep neural networks. *CoRR*, abs/1902.06531, 2019.

[7] David Ha, Andrew M. Dai, and Quoc V. Le. Hypernetworks. *CoRR*, abs/1609.09106, 2016.

[8] Surya Mattu Julia Angwin, Jeff Larson and Lauren Kirchner. Machine bias. *International Journal of Science Education*, 2016.

[9] Ronny Kohavi and Barry Becker. UCI machine learning repository, 1994.

[10] Phung Lai, NhatHai Phan, Abdallah Khreishah, Issa Khalil, and Xintao Wu. Model transferring attacks to backdoor hypernetwork in personalized federated learning. *CoRR*, abs/2201.07063, 2022.

[11] Jaeho Lee, Sejun Park, Sangwoo Mo, Sungsoo Ahn, and Jinwoo Shin. A deeper look at the layerwise sparsity of magnitude-based pruning. *CoRR*, abs/2010.07611, 2020.

[12] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020.

[13] Xingyu Li, Zhe Qu, Shangqing Zhao, Bo Tang, Zhuo Lu, and Yao Liu. Lomar: A local defense against poisoning attack on federated learning. *IEEE Transactions on Dependable and Secure Computing*, pages 1–1, 2021.

[14] Paul Pu Liang, Terrance Liu, Ziyin Liu, Ruslan Salakhutdinov, and Louis-Philippe Morency. Think locally, act globally: Federated learning with local and global representations. *CoRR*, abs/2001.01523, 2020.

[15] Junjie Liu, Zhe Xu, Runbin Shi, Ray C. C. Cheung, and Hayden Kwok-Hay So. Dynamic sparse training: Find efficient sparse network from scratch with trainable masked layers. *CoRR*, abs/2005.06870, 2020.

[16] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[17] H. Brendan McMahan, Eider Moore, Daniel Ramage, and Blaise Agüera y Arcas. Federated learning of deep networks using model averaging. *CoRR*, abs/1602.05629, 2016.

[18] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5389–5400. PMLR, 09–15 Jun 2019.

[19] Atri Rudra. Compas dataset. *International Journal of Science Education*, 2020.

[20] Aviv Shamsian, Aviv Navon, Ethan Fetaya, and Gal Chechik. Personalized federated learning using hypernetworks. *CoRR*, abs/2103.04628, 2021.

[21] Samuil Stoychev and Hatice Gunes. The effect of model compression on fairness in facial expression recognition. *CoRR*, abs/2201.01709, 2022.

[22] Ziteng Sun, Peter Kairouz, Ananda Theertha Suresh, and H. Brendan McMahan. Can you really backdoor federated learning? *CoRR*, abs/1911.07963, 2019.

[23] Minh-Hao Van, Wei Du, Xintao Wu, and Aidong Lu. Poisoning attacks on fair machine learning. *CoRR*, abs/2110.08932, 2021.

[24] Mingyang Wan, Daochen Zha, Ninghao Liu, and Na Zou. Modeling techniques for machine learning fairness: A survey. *CoRR*, abs/2111.03015, 2021.

[25] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y. Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 707–723, 2019.

[26] Tian Xu, Jennifer White, Sinan Kalkan, and Hatice Gunes. Investigating bias and fairness in facial expression recognition, 2020.