# Investigating Trade-offs in Utility, Fairness and Differential Privacy in Neural Networks

Pannekoek, Marlotte; Spigler, Giacomo

*Published in:*
arXiv

*Publication date:*
2021

*Document Version*
Early version, also known as pre-print

# Investigating Trade-offs in Utility, Fairness and Differential Privacy in Neural Networks

**Marlotte Pannekoek** [1] **Giacomo Spigler** [1]

## Abstract

To enable an ethical and legal use of machine learning algorithms, they must both be fair and protect the privacy of those whose data are being used. However, implementing privacy and fairness constraints might come at the cost of utility (Jayaraman & Evans, 2019; Gong et al., 2020). This paper investigates the privacy-utility-fairness trade-off in neural networks by comparing a Simple (S-NN), a Fair (F-NN), a Differentially Private (DP-NN), and a Differentially Private and Fair Neural Network (DPF-NN) to evaluate differences in performance on metrics for privacy ($\epsilon$, $\delta$), fairness (risk difference), and utility (accuracy). In the scenario with the highest considered privacy guarantees ($\epsilon = 0.1$, $\delta = 0.00001$), the DPF-NN was found to achieve better risk difference than all the other neural networks with only a marginally lower accuracy than the S-NN and DP-NN. This model is considered fair as it achieved a risk difference below the strict (0.05) and lenient (0.1) thresholds. However, while the accuracy of the proposed model improved on previous work from Xu, Yuan and Wu (2019), the risk difference was found to be worse.

## 1. Introduction

Machine learning algorithms are employed with great goals in mind, such as improved decision-making or increased efficiency. These benefits explain the ubiquitous use of such algorithms. However, potential harms should not be overlooked. Firstly, using personal data for training can cause leakage of private information. In addition, individuals can be unfairly affected when outcomes are dependent on race, sex, religious beliefs, sexual orientation, or economic status. For machine learning algorithms trained on personal data to be viable in today's society, they must ensure fairness and privacy for their users. This is required for both ethical and legal reasons (Xu et al., 2019; Wood et al., 2018). In this paper, we build on existing research to investigate how algorithms can ensure privacy while remaining fair and useful. To this extent, we tested how applying the fairness method Reject Option Classification influences the performance of a neural network (Briggs & Hollmén, 2020). Furthermore, the effects of adding a privacy-preserving optimizer to a simple and to a fair neural network were investigated.

Prior research has shown that when personal data are used for training machine learning models, these data can be retrieved by observing the behavior or structure of the learning algorithms (Yeom et al., 2020). This can negatively affect social status, employment chances, and insurance costs for end users of the systems (Wood et al., 2018). Therefore, it appears critical to develop and apply privacy-preserving methods. Along this direction, here we focus on the differential privacy framework, which is currently considered state-of-the-art (Gong et al., 2020). Its goal is to produce an approximately equal output, whether an individual is included in the analysis or not (Wood et al., 2018). An essential element of differential privacy is the privacy budget $\epsilon$, which controls how well privacy is protected. In general, lower values of $\epsilon$ imply more privacy protection. Nonetheless, choosing the best value for the parameter is still difficult (Jayaraman & Evans, 2019).

Finally, fairness entails that the behavior of a machine learning system does not dependent on protected attributes such as race and gender (Zhu et al., 2020). It is often assumed that algorithms are fair and impartial since decisions are based on data instead of human judgment (Corbett-Davies & Goel, 2018). However, several studies provide evidence that algorithms can be negatively biased towards protected groups, which could, for example, cause men to be preferred over women with similar skills by hiring algorithms (Xu et al., 2019). Several methods have been developed to achieve fairness via manipulation of the data or of the algorithms during pre-processing, in-processing, or post-processing phases. In this research, the post-processing method of "Reject Option Classification" was applied to a simple and a differentially private neural

[1]Department of Cognitive Science and Artificial Intelligence, Tilburg University, Tilburg, Netherlands. Correspondence to: Giacomo Spigler <g.spigler@uvt.nl>.

network to explore its effect on the performance of the network.

The existing empirical research on differentially private and fair algorithms has mainly focused on logistic regression (Xu et al., 2019; Jagielski et al., 2019; Ding et al., 2020). Research by Friedler et al. (2019), however, shows that the trade-off between accuracy and fairness can differ greatly depending on the specific algorithm being used. This suggests that it would be useful to research the combination of fairness and privacy on a wider range of machine learning systems.

In recent years both differential privacy and fairness have received much attention (Caton & Haas, 2020; Yeom et al., 2020). However, several authors have argued more research on algorithms that can achieve both goals (Ekstrand et al., 2018; Zhu et al., 2020; Xu et al., 2019; Datta et al., 2018). Earlier work on differential privacy has shown that there can be trade-offs between utility and privacy-preservation (Jayaraman & Evans, 2019; Gong et al., 2020). Similar work has been done on the trade-off between utility and fairness (Feldman et al., 2015). *Here we investigate the trade-off between utility, privacy, and fairness when the goal of an algorithm is to perform well on all three metrics.*

Specifically, we explore how the privacy-utility-fairness trade-off in neural networks is affected by:
- the Reject Option Classification method for fairness;
- a differentially private optimizer, under varying privacy budgets;
- the joint use of a differentially private optimizer and the Reject Option Classification method for fairness.

We finally compare the different cases with corresponding state-of-the-art results by Xu, Yuan, and Wu (2019).

## 2. Related Work

The combination of fairness and differential privacy has been the subject of recent empirical research (Xu et al., 2019). The work by Xu, Yuan and Wu (2019) in particular compared two algorithms with respect to utility and fairness, under different choices of the privacy budget ($\epsilon$). The algorithms were tested on two benchmarks; the Adult and the Dutch dataset. The metrics used for utility and fairness were accuracy and risk difference, respectively. For their first algorithm, logistic regression was used, and a penalty was added to the objective function to ensure fairness. A technique known as the functional mechanism was then applied to the objective function to ensure differential privacy. In the second algorithm, the objective function was corrupted by noise sampled from a Laplace distribution of which the mean was shifted according to a fairness constraint, therefore satisfying both fairness and differential privacy. Both algorithms achieved differential privacy and fairness with reasonable utility. Their results show that as $\epsilon$ was decreased, thus increasing the degree of privacy protection, the accuracy decreased, with a dynamics consistent across both datasets and methods.

Further, Jagielski et al. (2019) compared the use of a post-processing method for achieving fairness with an in-processing method. Only the algorithm using post-processing achieved a reasonable fairness-accuracy-privacy trade-off. However, only privacy with regard to the sensitive attribute was taken into account.

Ding et al. (2020) also combined privacy and fairness and explored varying the amount of noise added to different attributes. They also used a less strict definition of differential privacy. Using these techniques, they improved on the results achieved by Xu, Yuan and Wu (2019). Their results suggest that increasing $\epsilon$ from 0.1 to 1 results in a higher increase in accuracy than increasing $\epsilon$ from 10 to 100. However, this difference was not tested for significance. For most values of $\epsilon$, their algorithms achieved reasonable values of risk difference.

Cummings, Gupta, Kimpara and Morgenstern (2019) found that exact fairness and differential privacy in their set-up could not be achieved, although a trade-off could be achieved by relaxing the notion of exact fairness to 'approximate fairness'.

Hajian, Domingo-Ferrer, Monreale, Pedreschi and Giannotti (2015) investigated the application of privacy-preserving and fairness techniques on patterns derived by a classifier using multiple association rules. Their approach can be classified as a post-processing approach as they altered the classifier's results instead of the classifier itself or the data it was trained on (Hajian et al., 2015). They investigate both $k$-anonymity and differential privacy to achieve privacy. Their empirical analyses demonstrate that $k$-anonymity distorted the patterns less than the differential privacy approach. Additionally, their results show that applying the privacy technique after the fairness technique can deteriorate the achieved fairness.

## 3. Experiments

### 3.1. Dataset Description

The experiments presented here were performed using the Adult dataset (Dua & Graff, 2017), due to its importance in research on privacy and fairness, and to allow for a direct comparison with state-of-the-art results (Xu et al., 2019).

The importance of the Adult dataset in the field is due to the presence of sensitive attributes that could be used for personal identification or potentially threaten the fairness

of models trained on it. The dataset is openly available and consists of 45,222 cases and 14 variables. In this research, only 'sex' was regarded as a sensitive attribute, and 'income' was regarded as the dependent variable. Income is treated as a binary variable, separating incomes of less than 50,000 (income = 0) and more than 50,000 (income = 1). The sex variable refers to biological sex with male and female as the possible values.

### 3.2. Data Pre-Processing

To ensure comparability, the same pre-processing steps were applied as in the study by Xu, Yuan and Wu (2019). Specifically, list-wise deletions were performed, dummy codes were used for the categorical variables, and continuous variables were normalized. The Adult dataset already contained a separate train and test set. The train dataset was further split up into train and validation sets. The train, validation, and test dataset constitute 53.4%, 13.3%, and 33.3% of the total amount of data, respectively.

Initial data exploration showed a class imbalance in the labels, with approximately 75% of the reported incomes being less than 50,000. Furthermore, data exploration showed that females were underrepresented, making up around 32% of the cases. Approximately 88% of females earned less than 50,000, against 69% of males, which motivated applying fairness constraints.

### 3.3. Models

Four models were compared to explore the effect of differential privacy and fairness methods on the privacy-utility-fairness trade-off: a baseline 'Simple' neural network (S-NN), and a Fair (F-NN), a Differentially Private (DP-NN), and a Differentially Private and Fair neural network (DPF-NN). All models were implemented using Keras (Chollet, 2015). When no specific parameter settings are mentioned, the default settings were used.

**Simple Neural Network (S-NN).** The S-NN consisted of three fully connected layers with six neurons in the first and second layer and one neuron in the final layer. The first and the second layer used a ReLu activation, while the last layer used a sigmoid activation. Binary cross-entropy was used as the loss function and Adam as the optimizer (Kingma & Ba, 2017). Training was performed for a fixed duration of 20 epochs of Stochastic Gradient Descent with minibatches of size $mb = 20$.

**Fair Neural Network (F-NN).** The network used for the F-NN was equal to the S-NN. However, the 'Reject Option Classification' method was added to alter the output labels after prediction in an effort to improve fairness. The method was chosen as the best performing of six fairness methods that were previously evaluated. Results from the comparison are reported in the Supplementary Materials. The pre-processing and post-processing techniques to ensure fairness were implemented using the Artificial Intelligence Fairness 360 library (AIF 360) (Bellamy et al., 2018).

**Differentially Private Neural Network (DP-NN).** The network used for the DP-NN was equal to the S-NN. However, training was performed using a differentially private variant of the Adam optimizer (DPAdamGaussianOptimizer) (McMahan et al., 2019). This optimizer adds Gaussian noise to the gradient to ensure differential privacy. The noise_multiplier parameter was used to specify the amount of noise added to the model. This parameter's value depends on the target value for $\epsilon$ and $\delta$ (quantifying the probability of not achieving privacy within the privacy budget.), and was calculated using the compute_dp_sgd_privacy function. Another notable aspect of the differentially private optimizer is that norm clipping is applied after the data are split up into minibatches, but before adding the noise. Training was repeated for values of $\epsilon \in \{0.1, 1, 10, 100\}$ and $\delta \in \{0.01, 0.001, 0.0001, 0.00001\}$, as they are commonly used values in differential privacy applications (Xu et al., 2019; Ding et al., 2020).

**Differentially Private and Fair Neural Network (DPF-NN).**

The DPF-NN finally integrated the F-NN and the DP-NN, combining the use of the Reject Option Classification method from F-NN with the differentially private optimizer from DP-NN. The same values of $\epsilon$ and $\delta$ as the DP-NN were used.

### 3.4. Evaluation Criteria

The different models were tested for accuracy and risk difference to assess the utility and fairness of the models, respectively. All experiments were repeated ten times with different random seeds. This procedure is compatible with the previous work by Xu, Yuan and Wu (2019).

A mean accuracy above 75.4% on test data was considered an improvement over an algorithm that chooses the majority label. Risk difference was used as the metric to evaluate model fairness. A risk difference of 0 was considered optimal for fairness. Standard thresholds below which models are considered fair include 0.05 and 0.1 (Briggs & Hollmén, 2020; Xu et al., 2019). Both were considered and will be referred to as the strict and the lenient threshold, respectively, throughout this paper.

The risk difference and accuracy scores (for given values of $\epsilon$ and $\delta$) were compared to the thresholds and the scores achieved by the baseline models by Xu, Yuan and Wu (2019). Independent t-tests were applied to assess significant differences between algorithms, using a significance

level of 0.05. Lastly, linear regression was performed to test for an effect of $\epsilon$ and $\delta$ on mean accuracy in the DP-NN and DPF-NN models.

## 4. Results

The accuracy and risk difference for all the neural networks ($\epsilon = 0.1$, $\delta = 0.00001$) are shown in Table 1, together with the results achieved by the equivalent models from Xu, Yuan and Wu (2019). Further comparisons between the models are available in the Supplementary Materials.

*Table 1.* Comparison between the performance of the neural networks from this work (top half) and the logistic regression models by Xu, Yuan and Wu (2019) (bottom half). The order of the models indicate correspondence between the models proposed in this work and those of Xu, Yuan and Wu (simple baseline - LR; fair network - FairLR; differentially private network - PrivLR; differentially private and fair network - PFLR*). The privacy parameters $\epsilon = 0.1$, $\delta = 0.00001$ were used in the models that applied differential privacy constraints. Performance is shown as mean accuracy (in percentage) and risk difference (with standard deviations). Higher values are preferred for accuracy, whereas lower values are preferred for risk difference.

| | ACCURACY | RISK DIFFERENCE |
|---|---|---|
| S-NN | 84.14 ± 0.34 | 0.1310 ± 0.0147 |
| DP-NN | 84.03 ± 0.05 | 0.1355 ± 0.0024 |
| F-NN | 79.25 ± 3.50 | 0.0566 ± 0.0065 |
| DPF-NN | 82.98 ± 0.19 | 0.0475 ± 0.0020 |
| LR | 83.80 ± 0.23 | 0.1577 ± 0.0064 |
| PRIVLR | 62.63 ± 14.80 | 0.0883 ± 0.0805 |
| FAIRLR | 77.39 ± 5.21 | 0.0095 ± 0.0071 |
| PFLR* | 74.91 ± 0.40 | 0.0028 ± 0.0039 |

T-tests were applied to assess whether the differences between the different neural networks and the models by Xu, Yuan and Wu (2019) are significant at a significance level of 0.05. The results from these t-tests regarding the mean accuracy and risk difference scores can be found in Table 2 and Table 3, respectively.

**Simple Neural Network (S-NN).** As shown in Table 1 the S-NN achieved an average accuracy of 84.14% (*SD* = 0.34). This is slightly but significantly higher than the simple logistic regression (LR) by Xu, Yuan and Wu (2019), $t(18) = 2.6$, $p = .017$, which achieved an average accuracy of 83.80% (*SD* = 0.23). The average risk difference for the S-NN was 0.1310 (*SD* = 0.0147), which is slightly but significantly lower than the 0.1577 (*SD* = 0.00064) risk difference from the model by Xu, Yuan and Wu (2019), $t(18) = -5.3$, $p < .001$. The achieved accuracy of the S-NN is above the threshold of an algorithm that chooses the majority label.

**Fair Neural Network (F-NN).**

The F-NN achieved an average accuracy of 79.25% (*SD* = 3.50), which is above the majority label threshold and a significant decrease of 4.89 compared to the S-NN (*M* = 84.14%, *SD* = 0.34), $t(18) = 4.4$, $p < .001$. Compared to the average accuracy from the fair model by Xu, Yuan and Wu (2019) (*M* = 77.39%, *SD* = 5.21), this is an increase of 0.0186 in average accuracy. However, this difference is not significant, $t(18) = 0.9$, $p = .361$. The mean risk difference achieved by the F-NN is 0.0566 (*SD* = 0.0065), which is slightly above the 0.05 threshold. Compared to the risk difference of the S-NN (*M* = 0.1310, *SD* = 0.0147), this is a 0.0744 decrease, which was found to be significant, $t(18) = -14.6$, $p < .001$. Compared to the fair logistic regression model by Xu, Yuan and Wu (2019), which achieved an average risk difference of 0.0095 (*SD* = 0.0071), this is a significant increase of 0.0471, $t(18) = 15.5$, $p < .001$.

In conclusion, the application of Reject Option Classification in the F-NN did lead to a decreased risk difference compared to the S-NN. The mean risk of the F-NN was below the lenient 0.1 threshold. However, it is still slightly above the 0.05 threshold and higher than the average risk difference achieved by the fair logistic regression model by Xu, Yuan and Wu (2019). The fair model's accuracy is lower compared to the S-NN but still acceptable and higher than that of the fair logistic model.

**Differentially Private Neural Network (DP-NN).**

A table can be found in the appendix that displays the mean accuracy and risk difference for the DP-NN with differing values for $\epsilon$ and $\delta$. A linear regression was run to determine whether there was a significant effect of $\epsilon$ and $\delta$ on average accuracy and risk difference for the DP-NN. No significant effect on risk difference ($F(6, 9) = 3.15, p = .0597, R^2 = 0.68$) or accuracy ($F(6, 9) = 2.88, p = .0748, R^2 = 0.66$) could be observed for varying values of $\epsilon$ and $\delta$. Across all $\delta$ and $\epsilon$ values, the overall average of the average accuracy is 84.05% (*SD* = 0.05). The overall average of the average risk difference is 0.1345 (*SD* = 0.0016).

The average accuracy of the DP-NN model with $\epsilon = 0.1$ and $\delta = 0.00001$, so with the highest privacy guarantee, is 84.03% (*SD* = 0.05). With the highest privacy guarantee, the DP-NN achieved a mean accuracy that was 0.11 lower than that of the S-NN (*M* = 84.14%, *SD* = 0.34). This difference was, however, not significant ($t(18) = -1.0$, $p = .325$). The average risk difference of the model with the highest privacy guarantee is 0.1355 (*SD* = 0.0024), a difference of 0.0045 compared to the S-NN (*M* = 0.1310, *SD* = 0.0147). However, this difference is not significant, $t(18) = 1.0$, $p = .352$.

In the appendix a summary table is given of the accuracy and risk difference for differing values of $\epsilon$ and $\delta = 0.00001$ for the DP-NN and the differentially private logistic regres-

*Table 2.* Difference in mean accuracy between all models. A positive score means that the model defined in the row performed better than the model defined in the column. Eq. Model refers to the equivalent model from Xu, Yuan and Wu (2019). For example, FairLR is the equivalent model for the F-NN. An independent t-test was performed to test if the difference in means was significant at significance level $\alpha = .05$ (indicated by *). $DF = 18$ for all t-tests. The t-statistic is reported in brackets. The privacy parameters $\epsilon = 0.1$, $\delta = 0.00001$ were used in the models that applied differential privacy constraints.

| | S-NN | DP-NN | F-NN | DPF-NN | EQ. MODEL |
|---|---|---|---|---|---|
| S-NN | 0 (0.0) | 0.11 (1.0) | 4.89* (4.4) | 1.16* (9.4) | 0.34* (2.6) |
| DP-NN | | 0 (0.0) | 4.78* (4.3) | 1.05* (16.9) | 21.40* (4.6) |
| F-NN | | | 0 (0.0) | -3.73* (-3.4) | 1.86 (0.9) |
| DPF-NN | | | | 0 (0.0) | 8.07* (57.6) |

*Table 3.* Difference in mean risk difference between all models. A negative score means that the model defined in the row performed better than the model defined in the column. Eq. Model refers to the equivalent model from Xu, Yuan and Wu (2019). For example, FairLR is the equivalent model for the F-NN.* means that the difference is significant at a 0.05 significance level. $DF = 18$ for all t-tests. The t-statistic is reported in brackets. The privacy parameters $\epsilon = 0.1$, $\delta = 0.00001$ were used in the models that applied differential privacy constraints.

| | S-NN | DP-NN | F-NN | DPF-NN | EQ. MODEL |
|---|---|---|---|---|---|
| S-NN | 0 (0.0) | -0.0045 (-1.0) | 0.0744* (14.6) | 0.0835* (17.8) | -0.0267* (-5.3) |
| DP-NN | | 0 (0.0) | 0.0789* (36.0) | 0.0880* (89.1) | 0.0472 (1.9) |
| F-NN | | | 0 (0.0) | 0.0091* (4.2) | 0.0471* (15.5) |
| DPF-NN | | | | 0 (0.0) | 0.0447* (32.3) |

sion model by Xu, Yuan and Wu (2019). When comparing the models with the lowest $\epsilon$, the DP-NN model's average risk difference ($M$ = 0.1355, $SD$ = 0.0024) is 0.0472 higher than that of the differentially private logistic regression ($M$ = 0.0883, $SD$ = 0.0805). However, this is not a significant difference, $t(18) = 1.9$, $p = .080$. The DP-NN model ($M$ = 84.03%, $SD$ = 0.05) does significantly improve the average accuracy by 21.40, $t(18) = 4.6$, $p < .001$, in comparison with the differentially private logistic regression ($M$ = 62.63%, $SD$ = 14.80). When comparing the models with the highest $\epsilon$, DP-NN also achieved a higher average accuracy ($M$ = 84.03%, $SD$ = 0.05) in comparison with the logistic regression model ($M$ = 82.95%, $SD$ = 0.32). The difference is smaller (10.08) but still significant, $t(18) = 10.5$, $p < .001$.

In conclusion, no trend in average accuracy or risk difference was found for varying values of $\epsilon$ and $\delta$. Furthermore, there were no significant differences in average accuracy or risk difference between the DP-NN and the S-NN ($\epsilon = 0.1$, $\delta = 0.00001$). The DP-NN does, however, improve the average accuracy compared to the differentially private logistic regression.

**Differentially Private and Fair Neural Network (DPF-NN).** In the appendix a table is provided that shows the average accuracy and risk difference for varying values of $\epsilon$ and $\delta$ for the DPF-NN. As with the DP-NN, the mean risk difference and the accuracy barely differ for different values of $\delta$ and $\epsilon$. This is supported by the results from a simple linear regression that was performed to determine whether there was a significant effect of $\epsilon$ and $\delta$ on mean accuracy and risk difference. These results show no significant effect on mean accuracy ($F(6, 9) = 0.66, p = .687, R^2 = 0.30$) or risk difference ($F(6, 9) = 0.57, p = .748, R^2 = 0.27$). Across all values of $\delta$ and $\epsilon$ the average of the mean accuracy scores is 82.96 % ($SD$ = 0.25). The overall average for the average risk difference is 0.0475 ($SD$ = 0.0017). All averages for risk difference, including the overall average, are below the 0.05 threshold.

The DPF-NN with the highest privacy guarantee ($\epsilon$ = 0.1 and $\delta$ = 0.00001) has a mean accuracy of 82.98% ($SD$ = 0.19). Compared to the S-NN ($M$ = 84.14%, $SD$ = 0.34), that is a 1.16 lower average accuracy. This is a significant difference ($t(18)$ = -9.4, $p < .001$). The DPF-NN with the highest privacy guarantee achieved a mean risk difference of 0.0475 ($SD$ = 0.0020), which is a 0.0835 lower average risk difference compared to the S-NN ($M$ = 0.1310, $SD$ = 0.0147). This is also a significant difference, $t(18)$ = -17.8, $p < .001$.

Compared to the F-NN ($M$ = 79.25%, $SD$ = 3.50), the most private DPF-NN ($M$ = 82.98%, $SD$ = 0.19) has a significantly higher average accuracy ($t(18)$ = 3.4, $p$ = .003). The difference between the models is 3.73. The average risk

difference of the DPF-NN ($M$ = 0.0475, $SD$ = 0.0020) is 0.0091 lower compared to the most private F-NN ($M$ = 0.0566, $SD$ = 0.0065), which is a significant difference ($t(18)$ = -4.2, $p$ = .001).

When $\epsilon$ equals 0.1 and $\delta$ equals 0.00001, the average accuracy of the DPF-NN ($M$ = 82.98%, $SD$ = 0.19) is significantly lower by 1.05 ($t(18)$ = -16.9, $p < .001$) compared to the DP-NN ($M$ = 84.03%, $SD$ = 0.05). With respect to average risk difference, the DPF-NN ($M$ = 0.0475, $SD$ = 0.0020) has a significantly lower average compared to the DP-NN ($M$ = 0.1355, $SD$ = 0.0024). This difference of 0.0880 is significant ($t(18)$ = -89.1, $p < .001$).

When the results for $\delta$ = 0.00001 for the DPF-NN are compared to the results from the PFLR* model by Xu, Yuan and Wu (2019), some conclusions can be drawn. Firstly, for all values of $\epsilon$, the PFLR* model achieves lower average risk difference than the DPF-NN model. However, the average accuracy scores are also lower for the PFLR*. The difference in risk difference between the two models is at most 0.0447, with the lowest $\epsilon$. In this case the difference between the DPF-NN ($M$ = 0.0475, $SD$ = 0.0020) and the PFLR* ($M$ = 0.0028, $SD$ = 0.0039) is significant ($t(18)$ = 32.3, $p < .001$). The minimal difference in risk difference between the DPF-NN ($M$ = 0.0437, $SD$ = 0.0154) and the PFLR* ($M$ = 0.0204, $SD$ = 0.0140) is 0.0233 when $\epsilon$ equals 10. This difference is still significant ($t(18)$ = 3.5, $p$ = .002).

For average accuracy, the largest difference between the DPF-NN ($M$ = 82.98%, $SD$ = 0.19) and the PFLR* ($M$ = 74.91%, $SD$ = 0.40) was 8.07, for the lowest privacy budget. This is a significant difference ($t(18)$ = 57.6, $p < .001$). The smallest difference in average accuracy between the DPF-NN ($M$ = 83.04%, $SD$ = 0.23) and the PFLR* ($M$ = 79.13%, $SD$ = 2.00) was for the highest $\epsilon$ and came down to a difference of 3.91. This smallest difference is also significant ($t(18)$ = 6.1, $p < .001$).

In conclusion, adding differential privacy and fairness to a simple neural network decreased the average risk difference below the 0.05 threshold. The average accuracy also decreased but was still well above the 75.4% threshold. Adding both differential privacy and fairness compared to only adding fairness increased the accuracy and decreased the risk difference. Adding both differential privacy and fairness compared to only adding differential privacy decreased the risk difference and slightly decreased the accuracy. When the DPF-NN ($\delta$ = 0.00001) is compared to the DPFLR* by Xu, Yuan and Wu (2019), the DPF-NN achieves a higher average risk difference, though still under the strict 0.05 threshold, but higher average accuracy for all values of $\epsilon$.

## 5. Discussion

Four models were compared, a Simple (S-NN), a Fair (F-NN), a Differentially Private (DP-NN), and a Differentially Private and Fair Neural Network (DPF-NN). These models were evaluated for different values of $\epsilon$ and $\delta$ on fairness and utility metrics. The models were compared relative to each other, to the models by Xu, Yuand and Wu (2019), and to several threshold values.

**Effects of Fairness Constraints on the Fairness-Utility Trade-off.**

Adding fairness constraints to the S-NN significantly reduced the mean accuracy but also significantly decreased the mean risk difference. The F-NN model achieved a mean risk difference of 0.0566 ($SD$ = 0.0065). This is below the lenient 0.1 threshold. However, it is close to but not below the strict 0.05 threshold. The achieved accuracy was above the threshold of choosing the majority label. Reductions in statistical parity also decreased utility when Reject Option Classification was applied in the research by Hufthammer, et al. (2020). However, in the research by Briggs and Hollmén (2020), the utility was improved.

**Effects of Privacy Constraints on the Privacy-Utility Trade-off.** The performance of the DP-NN was evaluated for different values of $\epsilon$ and $\delta$. A linear regression showed that no significant effect on risk differences or accuracy could be observed as a function of $\epsilon$ and $\delta$. Furthermore, changing the standard Adam optimizer of the S-NN to the differentially private optimizer in the DP-NN did not significantly change the accuracy nor the risk difference when the highest privacy constraints were applied ($\epsilon$ = 0.1, $\delta$ = 0.00001). This model achieved an accuracy well above the majority label baseline. The risk difference was above both the strict and the lenient thresholds, which is expected since no fairness constraints were added to this model. *Most notably and contrary to previous results (Zhao et al., 2020; Jayaraman & Evans, 2019; Gong et al., 2020), no disruption in accuracy was observed when using lower values of $\epsilon$ and $\delta$.*

**Effects of Fairness and Privacy Constraints on the Privacy-Utility-Fairness Trade-off.** Like the case of DP-NN, no trend in risk differences or accuracy was observed for varying values of $\epsilon$ and $\delta$ in the DPF-NN model. The results on the DPF-NN showed that adding both fairness and a strong privacy guarantee ($\epsilon$ = 0.1, $\delta$ = 0.00001) to the S-NN model significantly increased its accuracy. However, this also significantly decreased risk difference. While the accuracy of the DPF-NN with $\epsilon$ = 0.1 and $\delta$ = 0.00001 was lower, it was still well above the majority label baseline. The achieved mean risk difference by this model was below both the strict and lenient thresholds. The model is, therefore, considered fair. Adding both fairness and differ-

ential privacy ($\epsilon = 0.1$, $\delta = 0.00001$) significantly increased accuracy compared to only adding fairness, but it decreased accuracy compared to only adding differential privacy ($\epsilon = 0.1$, $\delta = 0.00001$). Adding both differential privacy and fairness improved fairness compared to adding only differential privacy or fairness.

**Comparison with Equivalent Models.** The results from Xu, yuan and Wu (2019) were used as baselines to compare the performance of the proposed models.

The S-NN achieved a significantly higher mean accuracy and a lower mean risk difference than the baseline simple model. The F-NN achieved higher accuracy than the baseline fair model though this difference is not significant. The F-NN also produced a significantly higher mean risk difference and is, therefore, deemed less fair. There were no significant changes in risk difference between the DP-NN with the highest privacy guarantees ($\epsilon = 0.1$, $\delta = 0.00001$) and the PFLR* with the highest privacy guarantees ($\epsilon = 0.1$) by Xu, Yuan and Wu (2019). The DP-NN did, however, achieve a significantly higher mean accuracy. The DPF-NN with the highest privacy guarantees ($\epsilon = 0.1$, $\delta = 0.00001$) significantly outperformed the baseline differentially private and fair model with the highest privacy guarantees ($\epsilon = 0.1$) on mean accuracy but produced a significantly higher mean risk difference. It is, thus, less fair but has more utility.

**Limitations and Future Research.**

In the present work, only the variable 'sex' was considered as a sensitive attribute. However, in many practical applications, multiple sensitive variables may need to be considered. For example, the models may be considered fair with respect to gender but still make discriminatory decisions with respect to race. Likewise, Caton and Haas (2020) also warned of the effects of variables that are themselves not considered sensitive, but are still related to sensitive variables. Furthermore, only one fairness metric was considered in this research, risk difference, which is a measure of demographic/statistical parity that captures group fairness. As previously discussed, reduced unfairness according to one metric may not reduce and it may even increase unfairness according to another metric (Lee & Kizilcec, 2020; Hufthammer et al., 2020; Caton & Haas, 2020). If the models used in this research would be employed in real-life settings it would, therefore, be crucial to consider if demographic/statistical parity would be suitable for the specific application. Related to this are the chosen threshold: for example, in this research, the limit of risk difference beneath which a model was deemed fair was 0.05 for the strict threshold and 0.1 for the lenient threshold. However, whether these thresholds are suitable in a real-life application may depend on the use case and legislative requirements (Datta et al., 2018). If the goal of fairness is preferred over utility, the fair and differentially private and fair logistic regression models by Xu, Yuan and Wu (2019) should be preferred over the equivalent models presented here. Likewise, utility may be measured differently in different applications, which may require using balanced accuracy, F1-score, or other metrics, which might lead to different model rankings. Additionally, it would be interesting to assess whether combining pre- and post- processing fairness techniques would improve results. Lastly, a relaxed notion of differential privacy was considered in this research, while in some applications, strict or traditional differential privacy may be preferred.

Lastly, the results achieved by the differentially private and fair model in this research are encouraging and should be validated on a larger selection of datasets and using other metrics to assess the privacy-utility-fairness trade-off.

## 6. Conclusion

In this paper, we explored the impact of differential privacy and fairness constraints on the privacy-utility-fairness, both when they are applied independently and when they are combined together. Contrary to the previous research on this topic, this research focused on neural networks instead of logistic regression. Applying only fairness constraints led to a model with high accuracy and fairness but no privacy. Applying only privacy constraints led to a private but unfair model with high accuracy. The model that combined privacy and fairness constraints achieved better fairness than the model that only applied fairness constraints, while maintaining high accuracy. While the accuracy of the fair and private model significantly improved on previous work from Xu, Yuan and Wu (2019), the risk difference was found to be worse. Contrary to previous research both the DP-NN and DPF-NN model did not show a trend of a decrease in accuracy with an increase in the offered privacy. In conclusion, creating models that achieve fairness and preserve privacy while maintaining satisfactory utility is both possible and necessary. Hopefully, this and other contributions to the existing research on private and fair models can encourage and improve their use in real-world applications.

## References

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martina, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., and Zhang, Y. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *https://arxiv.org/abs/1810.01943*, October 2018.

Briggs, E. and Hollmén, J. Mitigating discrimination in clinical machine learning decision support using algorithmic processing techniques. In *International Conference on Discovery Science*, pp. 19–33. Springer, 2020.

Caton, S. and Haas, C. Fairness in Machine Learning: A Survey. *arXiv:2010.04053 [cs, stat]*, October 2020. arXiv: 2010.04053.

Chollet, F. Keras, 2015. URL https://github.com/fchollet/keras.

Corbett-Davies, S. and Goel, S. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. *arXiv:1808.00023 [cs]*, August 2018. URL http://arxiv.org/abs/1808.00023. arXiv: 1808.00023.

Cummings, R., Gupta, V., Kimpara, D., and Morgenstern, J. On the Compatibility of Privacy and Fairness. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization - UMAP'19 Adjunct*, pp. 309–315, Larnaca, Cyprus, 2019. ACM Press. ISBN 978-1-4503-6711-0. doi: 10.1145/3314183.3323847. URL http://dl.acm.org/citation.cfm?doid=3314183.3323847.

Datta, A., Sen, S., and Tschantz, M. C. Correspondences between Privacy and Nondiscrimination: Why They Should Be Studied Together. *arXiv:1808.01735 [cs]*, August 2018. arXiv: 1808.01735.

Ding, J., Zhang, X., Li, X., Wang, J., Yu, R., and Pan, M. Differentially private and fair classification via calibrated functional mechanism. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 622–629, 2020. Issue: 01.

Dua, D. and Graff, C. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Ekstrand, M. D., Joshaghani, R., and Mehrpouyan, H. Privacy for All: Ensuring Fair and Equitable Privacy Protections. In *Proceedings of Machine learning Research: Conference on Fairness, Accountability, and Transparency*, pp. 1–13, 2018.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15*, pp. 259–268, Sydney, NSW, Australia, 2015. ACM Press. ISBN 978-1-4503-3664-2. doi: 10.1145/2783258.2783311. URL http://dl.acm.org/citation.cfm?doid=2783258.2783311.

Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., and Roth, D. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pp. 329–338, New York, NY, USA, January 2019. Association for Computing Machinery. ISBN 978-1-4503-6125-5. doi: 10.1145/3287560.3287589. URL https://doi.org/10.1145/3287560.3287589.

Gong, M., Pan, K., Xie, Y., Qin, A., and Tang, Z. Preserving differential privacy in deep neural networks with relevance-based adaptive noise imposition. *Neural Networks*, 125:131–141, May 2020. ISSN 08936080. doi: 10.1016/j.neunet.2020.02.001. URL https://linkinghub.elsevier.com/retrieve/pii/S089

Hajian, S., Domingo-Ferrer, J., Monreale, A., Pedreschi, D., and Giannotti, F. Discrimination- and privacy-aware patterns. *Data Mining and Knowledge Discovery*, 29(6):1733–1782, November 2015. ISSN 1573-756X. doi: 10.1007/s10618-014-0393-7. URL https://doi.org/10.1007/s10618-014-0393-7.

Hufthammer, K. T., Aasheim, T. H., Ånneland, S., Brynjulfsen, H., and Slavkovik, M. Bias mitigation with aif360: A comparative study. In *Norsk IKT-konferanse for forskning og utdanning*, number 1, 2020.

Jagielski, M., Kearns, M., Mao, J., Oprea, A., Roth, A., Sharifi-Malvajerdi, S., and Ullman, J. Differentially Private Fair Learning. In *International Conference on Machine Learning*, pp. 3000–3008, 2019.

Jayaraman, B. and Evans, D. Evaluating differentially private machine learning in practice. In *28th {USENIX} Security Symposium ({USENIX} Security 19)*, pp. 1895–1912, 2019.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization, 2017.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Lee, H. and Kizilcec, R. F. Evaluation of Fairness Trade-offs in Predicting Student Success. *arXiv:2007.00088 [cs]*, June 2020. arXiv: 2007.00088.

McMahan, H. B., Andrew, G., Erlingsson, U., Chien, S., Mironov, I., Papernot, N., and Kairouz, P. A General Approach to Adding Differential Privacy to Iterative Training Procedures. *arXiv:1812.06210 [cs, stat]*, March 2019. URL http://arxiv.org/abs/1812.06210. arXiv: 1812.06210.

Wood, A., Altman, M., Bembenek, A., Bun, M., Gaboardi, M., Honaker, J., Nissim, K., O'Brien, D., Steinke, T., and Vadhan, S. Differential Privacy: A Primer for a Non-Technical Audience. *SSRN Electronic Journal*, 2018. ISSN 1556-5068. doi: 10.2139/ssrn.3338027. URL https://www.ssrn.com/abstract=3338027.

Xu, D., Yuan, S., and Wu, X. Achieving differential privacy and fairness in logistic regression. In *Companion Proceedings of The 2019 World Wide Web Conference*, pp. 594–599, 2019.

Yeom, S., Giacomelli, I., Menaged, A., Fredrikson, M., and Jha, S. Overfitting, robustness, and malicious algorithms: A study of potential causes of privacy risk in machine learning. *Journal of Computer Security*, 28(1):35–70, February 2020. ISSN 18758924, 0926227X. doi: 10.3233/JCS-191362. URL https://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/JCS-191362.

Zhao, B. Z. H., Kaafar, M. A., and Kourtellis, N. Not one but many Tradeoffs: Privacy Vs. Utility in Differentially Private Machine Learning. In *Proceedings of the 2020 ACM SIGSAC Conference on Cloud Computing Security Workshop*, pp. 15–26, Virtual Event USA, November 2020. ACM. ISBN 978-1-4503-8084-3. doi: 10.1145/3411495.3421352. URL https://dl.acm.org/doi/10.1145/3411495.3421352.

Zhu, T., Ye, D., Wang, W., Zhou, W., and Yu, P. S. More than privacy: applying differential privacy in key areas of artificial intelligence. *arXiv preprint arXiv:2008.01916*, 2020.