Discovering and Understanding Algorithmic Biases in Autonomous Pedestrian Trajectory Predictions

Andrew Bae
andrew.bae@stonybrook.edu
Stony Brook University
Stony Brook, New York, USA

Susu Xu
susu.xu@stonybrook.edu
Stony Brook University
Stony Brook, New York, USA

ABSTRACT

Pedestrian trajectory prediction is an important module in autonomous vehicles (AVs) to ensure safe and effective motion planning. Recently, many deep learning algorithms that achieve near real-time trajectory predictions have been developed. However, people in the artificial intelligence (AI) ethics community have raised critical concerns about the bias and fairness of many general deep learning algorithms. For example, most pedestrian trajectory data is collected from majority populations, and models learned from this data may not generalize well to the heterogeneous needs and behavior patterns of different pedestrian groups, especially for vulnerable pedestrians like the disabled, the elderly, and children. Biases present in trajectory prediction algorithms could mean that pedestrians from certain vulnerable demographics are more likely to be involved in vehicle crashes. In this work, we test two state-of-the-art pedestrian trajectory prediction models for age and gender biases across three different datasets. We design and utilize novel evaluation metrics for comparing model performance. We find that both models perform worse on children and the elderly compared to adults. However, their performance is similar between men and women. We identify potential sources of these biases, as well as discuss several limitations of our study. Our future work will consist of testing more models, refining our evaluation metrics, further differentiating the dataset bias from the algorithmic bias, and mitigating the algorithmic biases.

CCS CONCEPTS

• **Social and professional topics** → **Age**; **Gender**; • **Computing methodologies** → **Machine learning**.

KEYWORDS

bias, fairness, trajectory prediction, algorithm evaluation

ACM Reference Format:
Andrew Bae and Susu Xu. 2022. Discovering and Understanding Algorithmic Biases in Autonomous Pedestrian Trajectory Predictions. In *The 20th ACM Conference on Embedded Networked Sensor Systems (SenSys '22), November 6–9, 2022, Boston, MA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3560905.3568433

SenSys '22, November 6–9, 2022, Boston, MA, USA

ACM ISBN 978-1-4503-9886-2/22/11…$15.00
https://doi.org/10.1145/3560905.3568433

1 INTRODUCTION

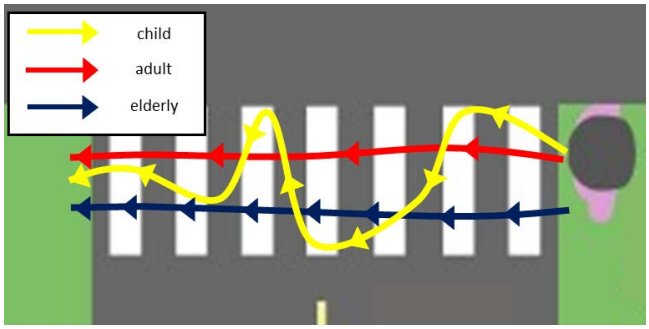

Figure 1: This figure is an exaggerated conceptual illustration that shows the differences in pedestrian walking patterns across different age groups.

Pedestrian safety remains a critical challenge in current transportation systems. In the US, fatal pedestrian crashes have increased by nearly 50% over the past decade [27]. However, experts in sustainable transportation argue that these traffic accidents are not really "accidents," as they stem from systemic inequalities ingrained in our society [30]. Data shows that children, the elderly, men, people with low income, the homeless, and people of color are involved in far greater pedestrian-vehicle crashes compared to the general population [1, 15, 20, 23].

AVs are expected to change the interaction between pedestrians and vehicles thanks to their advanced and precise control technology. In recent years, various pedestrian-vehicle collision prevention systems have been proposed by detecting pedestrians and predicting future trajectories using data sensed from on-board cameras and machine learning algorithms [24, 36]. With these technologies, AVs can effectively detect pedestrians and react to potential accidents. For example, Waymo claims that the cameras in their 5th-generation Waymo Driver can recognize pedestrians and signposts up to 500 meters away [16]. These pedestrian perception algorithms heavily rely on sensed data and high-capacity models to learn complex pedestrian behavior patterns.

Bias and lack of fairness are critical problems that exist in many AI systems today [2, 22, 31, 35]. With the development of the Internet of Things, various cameras and smart devices are deployed on vehicles to capture the data from the surrounding environment [7, 38, 39]. However, due to the constrained mobilities of vulnerable road users, data from vulnerable pedestrians, such as the elderly and children, is often limited. These vulnerable pedestrians have distinct distribution patterns compared to other pedestrian

groups [9]. For example, children are more likely to exhibit unpredictable behaviors [10], and elderly pedestrians on average walk slower than the general population [19], as shown conceptually in Figure 1. Additionally, men tend to display more risky and impulsive behaviors than women [14]. The data scarcity and distinct distributions of these underrepresented and disadvantaged pedestrian groups will make their data minor "mode" or even "out-of-distribution" compared to the huge amount of training data from other pedestrian groups. This will lead to larger prediction errors for disadvantaged groups during the testing stage. Error-prone detection and trajectory prediction of vulnerable pedestrians may cause discriminatory decision-making against these groups, compromising their safety.

Previous works exploring biases in AV perception primarily focused on pedestrian detection. Hirota et al. [3] and Kogure et al. [17] found that state-of-the-art models had lower detection rates for children compared adults. Wilson et al. [37] found that model detection rates were higher for lighter-skinned people compared to darker-skinned people. To our knowledge, we are the first to investigate biases specifically in pedestrian trajectory prediction.

In this work, we explore biases in pedestrian trajectory prediction algorithms. Our contributions in this work can be summarized as follows:

- We design a pipeline for evaluating system fairness in pedestrian trajectory predictions and provide insights into the discriminatory system behaviors.
- We design and utilize novel evaluation metrics for quantifying the system fairness of general pedestrian trajectory prediction systems across different pedestrian groups, by differentiating the dataset bias and algorithmic bias.
- We test our metrics on state-of-the-art models using multiple widely utilized datasets. We find that state-of-the-art models perform worse on child and elderly pedestrians compared to adults. However, we found no clear disparity between men and women.

In Section 2, we introduce our datasets, models, fairness metrics, and evaluation pipeline. In Section 3, we compare and analyze the performance of the models across different demographics, discuss the limitations of our study, and plan our future work.

2 METHODS

In this section, we begin by giving a general overview of trajectory prediction and algorithmic fairness. Next, we introduce our datasets and models. Lastly, we explain our fairness metrics and evaluation pipeline.

2.1 Overview

2.1.1 Trajectory Prediction. Predicting future trajectories equips AVs with the necessary information to plan safe paths through complicated and interactive environments, avoiding crashes or near collisions. However, trajectories can be difficult to analyze due to various variables that influence pedestrians in real-time. In the past, researchers tried modeling pedestrian behavior based on simple rules and mechanics [4, 13]. However, the rise of deep learning has led to a transition from physics-based models to data-driven models. Data for trajectory prediction is usually in the form of video, and it is often collected from an on-board camera or from a bird's eye point of view. The frames of relevant video clips are then annotated with bounding boxes surrounding the pedestrians. At time t, the trajectories in the past τ frames can be represented as $X_t = [x_{t-\tau+1}, x_{t-\tau+2}, ..., x_t]$, where x_t is the bounding box coordinates of a pedestrian. The goal of trajectory prediction is to predict the future bounding box coordinates $Y_t = [y_{t+1}, y_{t+2}, ..., y_{t+\delta}]$ in the next δ frames.

2.1.2 Algorithmic Fairness. Algorithmic fairness is a topic that has sparked the interest of the AI community recently. However, there is still no universal definition of fairness that is applicable to every situation. Out of the many prominent definitions of fairness, the one we are interested in for this work is statistical parity (also known as demographic parity). Statistical parity is satisfied under the following condition:

$$P(\hat{Y}|A = a) = P(\hat{Y}|A = b) \tag{1}$$

where $\hat{Y}$ is the predictor and a, b are different groups. The probability of a positive outcome should be the same regardless of whether the person is in a protected or unprotected group [33]. Statistical parity in the context of our work would mean that the trajectory prediction accuracy is the same across all pedestrian demographics.

2.2 Datasets

We use three datasets in our study, each of them captured through a single on-board camera: Joint Attention in Autonomous Driving (JAAD) [18], Pedestrian Intention Estimation (PIE) [26], and Trajectory Inference using Targeted Action priors Network (TITAN) [21]. The JAAD dataset was filmed mostly in Kremenchuk, Ukraine, with some other filming done across cities in Canada, Germany, the US, and Ukraine. It is annotated at 30 Hz and contains a total of 2580 pedestrians. However, only pedestrians that were close to the on-board camera (648 pedestrians, roughly 25% of all pedestrians) have age and gender labels. The PIE dataset was filmed entirely in Toronto, Canada, and it is annotated at 30 Hz. It contains a total of 1835 pedestrians, all of which have age and gender labels. The TITAN dataset was filmed entirely in Tokyo, Japan, and it is annotated at 10 Hz. It contains a total of 8588 pedestrians, and all of them have age labels but no gender labels. We specifically chose these datasets because to our knowledge, they are the only major datasets designed for trajectory prediction that have any pedestrian demographic labels.

2.3 Models

We look at two state-of-the-art models in this work, BiTraP [40] and SGNet [34]. Both of these models have several variations. BiTraP is a goal-conditioned bi-directional multi-modal trajectory prediction model based on a conditional variational autoencoder. BiTraP has a deterministic version, BiTraP-D, and two multimodal versions, BiTraP-NP and BiTraP-GMM. SGNet also predicts future trajectories based on goals, but instead of using a single, long-term goal like BiTraP, it uses a stepwise goal estimator that predicts successive goals in the future. SGNet has a deterministic version, SGNet, and a multimodal version, SGNet$_{CVAE}$. We specifically chose these two models because they are currently the two best-performing models on JAAD and PIE, and they have open-source code available.

2.3.1 Deterministic Models. Deterministic models predict one single trajectory based on its observations. We look at BiTraP-D and SGNet in this work. In the future, we also plan on testing PIE$_{traj}$ [26].

2.3.2 Multimodal Models. Although there is one single trajectory that a pedestrian ends up taking, there are multiple possible trajectories that a pedestrian may take at any given time. This is why multimodal models have been gaining popularity recently. Multimodal models predict multiple possible trajectories based on their observations. In this work, we look at BiTraP-NP. In the future, we also plan on testing BiTraP-GMM and SGNet$_{CVAE}$.

2.4 Fairness Quantification Metrics

We define a "track" to be the bounding box coordinates of a pedestrian over a certain time interval. Each track has a length of two seconds and consists of two components: the first 0.5 s of the track are for observing, and the next 1.5 s are for predicting.

The trajectory evaluation metric we use for evaluating the accuracy of a prediction at a single frame is the mean squared error (MSE) of the bounding box coordinates of the pedestrian. MSE is the standard trajectory evaluation metric for many pedestrian trajectory prediction models.

Fairness evaluation metrics quantify the difference in a model's performance over different demographic groups. We design and utilize three different fairness evaluation metrics: the mean MSE, the Mann-Whitney U Test, and the Wasserstein distance.

Mean MSE. Traditionally, model performance on the JAAD and PIE datasets has been measured using three variations of the mean MSE across all tracks: 1) mean of the bounding box MSE averaged over the first 0.5 s, 1.0 s, and 1.5 s, 2) mean of the bounding box center MSE (C$_{MSE}$) averaged over 1.5 s, and 3) bounding box center final MSE (CF$_{MSE}$) at 1.5 s. However, the MSE error distributions are highly skewed right as shown in Figure 2. The means of these error distributions are heavily influenced by outliers, and thus they may not be reflective of a model's true performance. As result, we propose two additional evaluation metrics.

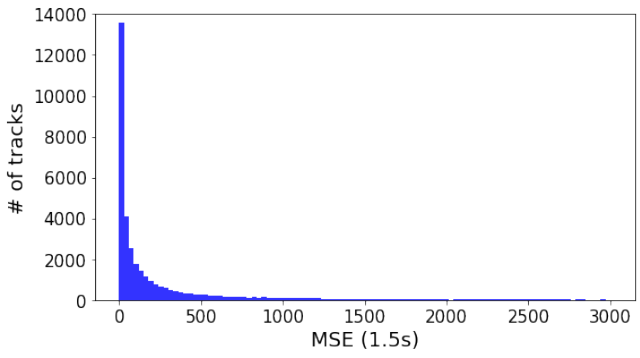

Figure 2: MSE (1.5s) error distribution of BiTraP-D performance on the PIE dataset. The skewed right trend is consistent for all models across all datasets. Note that some MSE (1.5s) values here go up to nearly 100,000, but we only plot up to 3000 in this figure.

Mann-Whitney U Test. The Mann-Whitney U Test is a non-parametric test that determines whether two independent samples derive from the same population. As stated before, the error distributions in our study are not normally distributed, so we cannot use a parametric test like the t-test. We conduct three different one-sided Mann-Whitney U Tests with the following null and alternative hypotheses:

- Children vs. Adults
  H_0: Model performance on children = adults
  H_1: Model performance on children < adults
- Elderly vs. Adults
  H_0: Model performance on the elderly = adults
  H_1: Model performance on the elderly < adults
- Men vs. Women
  H_0: Model performance on men = women
  H_0: Model performance on men < women

Wasserstein Distance. The Wasserstein distance is a distance function between distributions. It is essentially the amount of "work" required to transform one distribution into another. The Wasserstein distance between two distributions u and v is defined as follows:

$$W(u,v) := \inf_{\pi \in \Gamma(u,v)} \int_{\mathbb{R}\times\mathbb{R}} |x-y| d\pi(x,y) \tag{2}$$

where $\Gamma(u,v)$ is the set of distributions on $\mathbb{R}\times\mathbb{R}$ whose marginals are u and v on the first and second factors respectively. Zhao [41] proves that if a regressor is individually fair, which means the regressor treats similar individuals similarly, the gap of the regressor's accuracy disparity across groups with different sensitive attributes can be exactly measured by the Wasserstein distance. Therefore, we utilize the Wasserstein distance between the error distributions of children and adults, elderly and adults, and men and women to measure unfairness in model predictions across different demographic groups.

2.5 Fairness Evaluation Pipeline

We walk through our pipeline for evaluating algorithmic fairness in this subsection. For each dataset, we generate trajectory tracks for the train, validation, and test splits using a track overlap ratio of 0.5. We train each model on each dataset separately using all pedestrians in the training tracks, regardless of demographic. The testing tracks are split into groups according to pedestrian demographic. We test our trained model on each demographic group separately. For deterministic models, testing on a group that contains n test tracks results in the vector $d = [d_1, d_2, ..., d_n]$, where d_n is the bounding box MSE on track n. For multimodal models, testing on a group that contains n test tracks results in the matrix $E = [e_1, e_2, ..., e_n]$, where $e_n = [e_{n1}, e_{n2}, ..., e_{n20}]$ is a vector containing the bounding box MSEs of 20 randomly sampled trajectory predictions on track n. Following standards set by [12, 28, 29], we use the best-of-20 approach for multimodal models by extracting the best prediction on each track, turning the matrix E in the vector $m = [m_1, m_2, ..., m_n]$, where m_n is the minimum MSE value from e_n. Compared to deterministic models, the MSE values for multimodal models tend to be much lower because we essentially cherry-pick the most accurate trajectory prediction out of 20 predictions generated.

Implementation Details. For evaluation on JAAD and PIE, we use the pretrained model checkpoints publicly provided by each model's respective authors. For evaluation on TITAN, we train each model from scratch using each model's default hyperparameters, as none of the models have previous benchmarks on TITAN.

3 RESULTS AND DISCUSSION

In this section, we first expose biases in our datasets by showing their demographic distributions. Then, we test state-of-the-art model performance using the evaluation metrics introduced in Section 2.4 and provide insight into our results. Finally, we discuss the limitations of our study and outline our future work.

3.1 Dataset Bias

Biased data often leads to biased algorithmic outcomes [22]. Finding good quality unbiased data is a common challenge in machine learning, and our work is no exception. All three of our datasets are biased towards adults. This is somewhat expected, as in most societies, the population of adults is greater than the population of children and the elderly. However, there is an even greater percentage of adults in our datasets than what is expected according to their respective filming locations. Table 1 shows the pedestrian age distributions for each dataset and compares them to their expected age distributions. Some reasons for this disparity may be that children are likely to be at school, and the elderly are less likely to take walking trips compared to younger age groups [23]. These reasons would decrease the probability that a child or elderly pedestrian would make an appearance in one of the dataset videos. There is probably annotation bias as well. Many pedestrians in the videos are far away from the on-board camera, making it difficult to see their features. In previous work, Wilson et al. [37] found its annotators to have inconsistencies while labeling pedestrians as light-skinned or dark-skinned. It makes sense that there would also be inconsistencies in labeling age, which is arguably more subjective than skin tone. The annotators could have labeled pedestrians as adults when they were in doubt.

Table 1: Age demographic breakdown of JAAD, PIE, and TITAN datasets. We define children to be ages 0-14, adults to be ages 15-64, and elderly to be ages 65 and above. The JAAD dataset was filmed in 5 different cities across Ukraine, Canada, Germany, and the US, but 80% of the clips were filmed in Kremenchuk, Ukraine. We could not find age demographic data for Kremenchuk specifically, so we report the expected age demographics of Ukraine as a country.

Dataset	Statistic	Children	Adults	Elderly
JAAD	# of pedestrians	47	509	92
	% of dataset	7.3%	78.5%	14.2%
	% expected [8]	15.1%	69.3%	15.6%
PIE	# of pedestrians	17	1640	185
	% of dataset	0.9%	89.0%	10.0%
	% expected [11]	14.2%	68.1%	17.6%
TITAN	# of pedestrians	116	7872	506
	% of dataset	1.4%	91.7%	5.9%
	% expected [6]	11.5%	65.7%	22.8%

Table 2: Gender demographic breakdown of the JAAD and PIE datasets. We expect around a 50% distribution for both genders.

Dataset	Statistic	Male	Female
JAAD	# of pedestrians	277	355
	% of dataset	43.8%	56.2%
PIE	# of pedestrians	976	866
	% of dataset	53.0%	47.0%

3.2 Algorithmic Bias

Algorithmic bias is the bias that is added purely by the algorithm itself without any bias in the data. In our case, we don't know how much of the bias we measure actually stems from the algorithms because the models are all trained on biased input data, as we showed in Section 3.1. We compare model performance between age and gender using the mean MSE in Table 3, the Mann-Whitney U test in Table 4, and the Wasserstein distance in Figure 3. In the future, we will try to create unbiased datasets to train the models.

3.2.1 Age Bias. Our results show convincing evidence that there is a disparity in state-of-the-art pedestrian trajectory prediction model performance between age groups. Models tend to perform worse on child and elderly pedestrians compared to adult pedestrians.

1) Results for BiTraP-D: The mean MSE is the highest on children over all three datasets. The mean MSE is higher on children compared to adults by an average of 23%, 217%, and 25% over JAAD, PIE, and TITAN, respectively. The Mann-Whitney U Test p-values comparing elderly and adults are all statistically significant (they are all well under 10^{-15}) over PIE and TITAN.

2) Results for SGNet: The mean MSE is the highest on children over JAAD and PIE and the highest on the elderly over TITAN. Over PIE, the mean MSE is on average 254% worse on children and 42% worse on the elderly, compared to adults. Like BiTraP-D, the SGNet Mann-Whitney U test p-values comparing the elderly and adults are all significant over PIE and TITAN.

3) Results for BiTraP-NP: The mean MSE is the highest on children over PIE and the highest on the elderly over TITAN, and p-values comparing adults and elderly are all significant over PIE and TITAN. The mean MSE over JAAD is actually the highest on adults. While we are unsure of the reasons behind this, it is important to remember that BiTraP-NP is a multimodal model, meaning that for each track, it cherry-picks the trajectory with the lowest MSE error out of 20 randomly generated trajectories. This would not be possible in a real-world situation, as autonomous vehicles will not know the true future trajectory of a pedestrian. We are looking into alternatives to the standard best-of-20 approach for multimodal models.

The majority of p-values comparing children and adults over PIE and TITAN are not significant even though there is a big difference between their respective mean MSE values. This is likely due to the small sample size of children in these datasets. As Table 1 shows, children only make up 0.9% of the pedestrians in PIE and 1.4% of the pedestrians in TITAN. However, one trend that we notice is

Table 3: Model performance on different demographics in terms of mean MSE/C_{MSE}/CF_{MSE}. The worst performing age group and gender are bold.

		JAAD			PIE			TITAN		
Method	Group	MSE 0.5s / 1.0s / 1.5s	C_{MSE} 1.5s	CF_{MSE} 1.5s	MSE 0.5s / 1.0s / 1.5s	C_{MSE} 1.5s	CF_{MSE} 1.5s	MSE 0.5s / 1.0s / 1.5s	C_{MSE} 1.5s	CF_{MSE} 1.5s
BiTraP-D [40]	Child	**185** / **848** / **2826**	**2722**	**10945**	**85** / **468** / **1596**	**1572**	**7291**	289 / 1216 / **4218**	**4106**	**17161**
	Adult	182 / 662 / 2025	1900	7566	38 / 152 / 490	462	1880	360 / 1134 / 3120	2931	10421
	Elderly	147 / 410 / 1134	1040	4005	67 / 234 / 673	617	2346	**387** / **1374** / 4004	3773	13667
	Male	**191** / 611 / 1775	1641	6552	**43** / 160 / 490	454	1844	- / - / -	-	-
	Female	171 / **661** / **2064**	**1953**	**7806**	39 / **162** / **537**	**512**	**2127**	- / - / -	-	-
SGNet [34]	Child	147 / **736** / **2582**	**2481**	**10405**	**111** / **520** / **1581**	**1553**	**5940**	253 / 1033 / 3507	3418	13998
	Adult	**155** / 583 / 1844	1725	7177	33 / 133 / 449	422	1821	369 / 1174 / 3117	2945	10083
	Elderly	139 / 374 / 995	905	3479	61 / 202 / 586	533	2144	**392** / **1411** / **4243**	**4018**	**14698**
	Male	**166** / 545 / 1620	1494	6120	**38** / **141** / 449	416	1757	- / - / -	-	-
	Female	146 / **583** / **1876**	**1768**	**7236**	33 / 138 / **478**	**455**	**1976**	- / - /-	-	-
BiTraP-NP [40]	Child	**73** / 157 / 360	273	881	**43** / **173** / **457**	**415**	**1184**	158 / 296 / 566	509	1293
	Adult	72 / **167** / **390**	**303**	**993**	17 / 40 / 93	70	223	**187** / **356** / 697	554	1292
	Elderly	63 / 110 / 208	142	360	33 / 73 / 166	120	407	162 / 351 / **760**	**586**	**1572**
	Male	**78** / **161** / 341	247	740	**19** / **46** / **105**	76	253	- / - / -	-	-
	Female	64 / 157 / **385**	**307**	**1030**	16 / 40 / 97	**78**	**255**	- / - / -	-	-

Table 4: One sided Mann-Whitney U Test p-values. Statistically significant p-values (p < 0.05) are in bold.

		JAAD	PIE	TITAN
Method	Demographics Compared	MSE 0.5s / 1.0s / 1.5s	MSE 0.5s / 1.0s / 1.5s	MSE 0.5s / 1.0s / 1.5s
BiTraP-D [40]	Child|Adult	0.64 / **0.028** / **0.002**	0.58 / 0.42 / 0.24	0.84 / 0.12 / **0.008**
	Elderly|Adult	0.64 / 1.00 / 1.00	**5e-78** / **2e-68** / **3e-55**	**2e-23** / **2e-18** / **9e-18**
	Male|Female	**0.01** / 0.05 / 0.22	0.48 / 0.40 /0.21	- / - / -
SGNet [34]	Child|Adult	0.73 / **0.02** / **0.002**	0.53 / 0.23 / 0.10	1.00 / 0.71 / 0.09
	Elderly|Adult	0.21 / 1.00 / 1.00	**9e-76** / **1e-65** / **3e-54**	**5e-27** / **9e-34** / **7e-20**
	Male|Female	**0.01** / **0.04** / 0.05	0.15 / 0.10 / 0.23	- / - / -
BiTraP-NP [40]	Child|Adult	0.49 / **0.02** / **0.002**	0.54 / 0.24 / 0.21	1.00 / 0.94 / 0.31
	Adult|Elderly	0.10 / 0.98 / 1.00	**8e-93** / **8e-88** / **5e-81**	**9e-23** / **1e-23** / **9e-24**
	Male|Female	**0.005** / **0.01** / **0.02**	0.28 / 0.27 / 0.11	- / - / -

that the p-values progressively decrease as the prediction horizon increases (and thus the prediction problem inherently gets harder).

One part of our results that does not support our claim is that all models perform very well on elderly pedestrians over the JAAD dataset. We hypothesized that this had to do with the fact that in JAAD, only pedestrians that were close to the on-board camera are labeled with age and gender information (roughly 25%). Since the unlabeled pedestrians are further away, they naturally have lower error values. We thought that training on the large amount of unlabeled pedestrians had a significant impact on the performance between labeled pedestrians groups. As a result, we retrained all models on JAAD using only the labeled pedestrians, but found that model performance on the elderly was still significantly better than the other two age groups. This unusually good performance on elderly pedestrians is not present in the JAAD training set. The particularly accurate predictions of elderly pedestrian trajectories in JAAD are likely the result of high levels of statistical noise in the dataset.

3.2.2 Gender Bias. Model performance between men and women is overall pretty similar. Although some of the Mann-Whitney U test p-values are statistically significant, the corresponding differences in the mean MSEs are very small. The MSE 1.5s % difference between men and women is less than 15% on JAAD and less than 10% on PIE for all models. Additionally, the Wasserstein distances between genders are much smaller than the Wasserstein distances between age groups, as shown in Figure 3. While historical data shows that

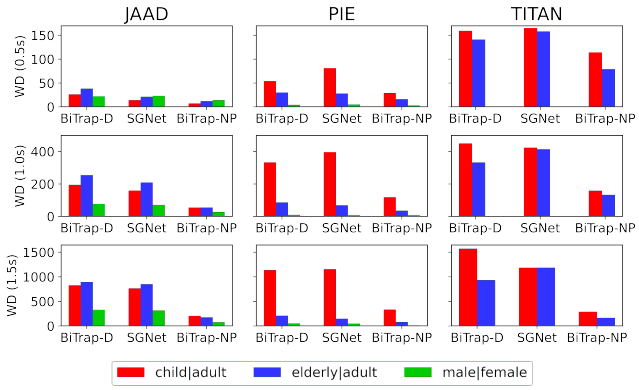

Figure 3: Wasserstein distances for the MSE error distributions between different demographics. It's important to note once again that on JAAD, the performance of all three models is better on elderley than on adults.

male pedestrians are involved in far more vehicle casualties than their female counterparts, the reasons behind this are not very clear. Zhu et al. [42] found that the biggest reason for this disparity was that men simply had greater fatality rates when involved in vehicle collisions. Tolea et al. [32] looked at the walking speeds of men and women, and found that while men on average walked slightly faster than women, the difference was not statistically significant. A better understanding of the reasons behind high male pedestrian casualties will be required to see if this disparity will persist with the widespread adoption of AVs.

3.3 Limitations

We have several limitations in our study. First off, we were unable to evaluate on other popular trajectory prediction datasets, such as ETH [25], UCY [19], or nuScenes [5], due to their lack of pedestrian demographic labels. We encourage new dataset makers to include pedestrian demographic information if possible, as potential biases in models trained using these datasets can lead to serious real-world consequences.

Additionally, the three datasets we use only have pedestrian labels for age and gender. Ideally, we would have also liked to study biases in other demographic factors as well, such as race or income. To our knowledge, no datasets designed for trajectory prediction label pedestrians with this information. However, we understand that this information may be difficult to obtain due to privacy concerns.

Also, the performance of all models on TITAN is overall pretty poor, as shown in Table 3. This is likely due to the fact that the TITAN dataset is only annotated at a frequency of 10 hz. With our current implementation details, models only have 5 frames to observe on TITAN (compared to 15 frames on both JAAD and PIE), making the prediction task much harder. Comparing model performance across different demographics may not give us much insight when the trajectory predictions are all fairly inaccurate. We tried transfer learning on TITAN using pretrained models from JAAD and PIE, but this did not have much effect on the final prediction accuracy.

Finally, the biggest limitation of our study is that the models are all trained on biased data, meaning that we cannot properly distinguish the dataset bias from the algorithmic bias. Creating unbiased datasets to train the models on will be necessary in order the gain a deeper understanding of the algorithmic biases.

3.4 Future Work

We have several future tasks planned ahead of us. Firstly, there are several models/variations of models that we have not yet tested (BiTraP-GMM, SGNet$_{CVAE}$, and PIE$_{traj}$). We plan on testing these models very shortly. Also, we will continue to refine our evaluation metrics. We are thinking of switching all MSE calculations to the mean absolute percentage error (MAPE) so that pedestrians close to the camera have similar error magnitudes compared to pedestrians further away. For multimodal models, we will look into alternatives to the standard best-of-20 approach. Additionally, we will attempt to create an unbiased dataset to train the models on so that we can properly distinguish the dataset bias from the algorithmic bias. Finally, once we gain a deeper understanding of the algorithmic biases, we will explore methods for mitigating these biases.

4 CONCLUSION

In this paper, we tested two state-of-the-art pedestrian trajectory prediction models for disparities in their performance between age and gender. In addition to the mean MSE, we also used the Mann-Whitney U test and the Wasserstein distance to compare performance across different demographics. We found that both models perform worse on children and the elderly compared to adults. We found no clear difference between genders. In future work, we will continue to test more models, refine our evaluation metrics, further differentiate the dataset bias and algorithmic bias, and explore methods for mitigating bias.

ACKNOWLEDGMENTS

This work was supported by the URECA (Undergraduate Research and Creative Activities) Program at Stony Brook University. We would like to thank Mohib Azam at Stony Brook University for helping set up several computer-related technologies used in this work.

REFERENCES

[1] National Highway Traffic Safety Administration. 2019. *Traffic Safety Facts 2017 Data - Pedestrians.* Technical Report. 216 pages. https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812681

[2] Shahriar Akter, Grace McCarthy, Shahriar Sajib, Katina Michael, Yogesh K. Dwivedi, John D'Ambra, and K. N. Shen. 2021. Algorithmic bias in data-driven innovation in the age of AI. *International Journal of Information Management* 60 (Oct. 2021), 102387. https://doi.org/10.1016/j.ijinfomgt.2021.102387

[3] Martim Brandao. 2019. Age and gender bias in pedestrian detection algorithms. https://doi.org/10.48550/arXiv.1906.10490 Number: arXiv:1906.10490 arXiv:1906.10490 [cs].

[4] C Burstedde, K Klauck, A Schadschneider, and J Zittartz. 2001. Simulation of pedestrian dynamics using a two-dimensional cellular automaton. *Physica A: Statistical Mechanics and its Applications* 295, 3 (June 2001), 507–525. https://doi.org/10.1016/S0378-4371(01)00141-8

[5] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. 2020. nuScenes: A Multimodal Dataset for Autonomous Driving. 11621–11631.

https://openaccess.thecvf.com/content_CVPR_2020/html/Caesar_nuScenes_A_Multimodal_Dataset_for_Autonomous_Driving_CVPR_2020_paper.html

[6] National Statistics Center. 2021. Population Census 2020 Population Census Basic Complete Tabulation on Population and HouseholdsPopulation, Households, Sex, Age and Marital status 2-3-2 Population composition ratio (by age) by Sex, Age (3 groups) and All nationality or Japanese - Japan, Prefectures, 21 Major Cities, Ku-area of Tokyo and Shi with population of 500,000 or more | View Statistical Table/Graph. https://www.e-stat.go.jp/en/dbview?sid=0003445136

[7] Xinlei Chen, Susu Xu, Jun Han, Haohao Fu, Xidong Pi, Carlee Joe-Wong, Yong Li, Lin Zhang, Hae Young Noh, and Pei Zhang. 2020. Pas: Prediction-based actuation system for city-scale ridesharing vehicular mobile crowdsensing. *IEEE Internet of Things Journal* 7, 5 (2020), 3719–3734.

[8] CIA. 2022. Ukraine. https://www.cia.gov/the-world-factbook/countries/ukraine/

[9] Piotr Czech. 2017. Physically disabled pedestrians—Road users in terms of road accidents. In *Contemporary challenges of transport systems and traffic engineering*. Springer, 157–165.

[10] Victoria Gitelman, Sharon Levi, Roby Carmel, Anna Korchatov, and Shalom Hakkert. 2019. Exploring patterns of child pedestrian behaviors at urban intersections. *Accident Analysis & Prevention* 122 (Jan. 2019), 36–47. https://doi.org/10.1016/j.aap.2018.09.031

[11] Statistics Canada Government of Canada. 2017. Census Profile, 2016 Census - Toronto, City [Census subdivision], Ontario and Toronto, Census division [Census division], Ontario. https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/prof/details/page.cfm?Lang=E&Geo1=CSD&Code1=3520005&Geo2=CD&Code2=3520&SearchText=toronto&SearchType=Begins&SearchPR=01&B1=All&TABID=1&type=0 Last Modified: 2019-06-18.

[12] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. 2018. Social GAN: Socially Acceptable Trajectories With Generative Adversarial Networks. 2255–2264. https://openaccess.thecvf.com/content_cvpr_2018/html/Gupta_Social_GAN_Socially_CVPR_2018_paper.html

[13] Dirk Helbing and Péter Molnár. 1995. Social force model for pedestrian dynamics. *Phys. Rev. E* 51, 5 (May 1995), 4282–4286. https://doi.org/10.1103/PhysRevE.51.4282 Publisher: American Physical Society.

[14] David Herrero-Fernández, Patricia Macía-Guerrero, Laura Silvano-Chaparro, Laura Merino, and Emily C. Jenchura. 2016. Risky behavior in young adult pedestrians: Personality determinants, correlates with risk perception, and gender differences. *Transportation Research Part F: Traffic Psychology and Behaviour* 36 (Jan. 2016), 14–24. https://doi.org/10.1016/j.trf.2015.11.007

[15] Kaci L. Hickox, Nancy Williams, Laurie F. Beck, Tom Coleman, John Fudenberg, Byron Robinson, and John Middaugh. 2014. Pedestrian Traffic Deaths Among Residents, Visitors, and Homeless Persons — Clark County, Nevada, 2008–2011. *MMWR Morb Mortal Wkly Rep* 63, 28 (July 2014), 597–602. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5779416/

[16] Satish Jeyachandran. 2020. Introducing the 5th-generation Waymo Driver: Informed by experience, designed for scale, engineered to tackle more environments. *Waymo LLC, březen* (2020).

[17] Shunsuke Kogure, Kai Watabe, Ryosuke Yamada, Yoshimitsu Aoki, Akio Nakamura, and Hirokatsu Kataoka. 2022. Age Should Not Matter: Towards More Accurate Pedestrian Detection via Self-Training. *Computer Sciences & Mathematics Forum* 3, 1 (2022), 11. https://doi.org/10.3390/cmsf2022003011 Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

[18] Iuliia Kotseruba, Amir Rasouli, and John K. Tsotsos. 2020. *Joint Attention in Autonomous Driving (JAAD)*. Technical Report arXiv:1609.04741. arXiv. https://doi.org/10.48550/arXiv.1609.04741 arXiv:1609.04741 [cs] type: article.

[19] Alon Lerner, Yiorgos Chrysanthou, and Dani Lischinski. 2007. Crowds by Example. *Computer Graphics Forum* 26, 3 (2007), 655–664. https://doi.org/10.1111/j.1467-8659.2007.01089.x _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-8659.2007.01089.x.

[20] Mike Maciag. 2014. Pedestrians Dying at Disproportionate Rates in America's Poorer Neighborhoods. https://www.governing.com/archive/gov-pedestrian-deaths-analysis.html Section: Archive.

[21] Srikanth Malla, Behzad Dariush, and Chiho Choi. 2020. TITAN: Future Forecast Using Action Priors. 11186–11196. https://openaccess.thecvf.com/content_CVPR_2020/html/Malla_TITAN_Future_Forecast_Using_Action_Priors_CVPR_2020_paper.html

[22] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.* 54, 6 (July 2021), 115:1–115:35. https://doi.org/10.1145/3457607

[23] Rebecca B. Naumann and Laurie F. Beck. 2013. Motor Vehicle Traffic-Related Pedestrian Deaths — United States, 2001–2010. *MMWR Morb Mortal Wkly Rep* 62, 15 (April 2013), 277–282. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4604973/

[24] Pedro J Navarro, Carlos Fernandez, Raul Borraz, and Diego Alonso. 2017. A machine learning approach to pedestrian detection for autonomous vehicles using high-definition 3D range data. *Sensors* 17, 1 (2017), 18.

[25] S. Pellegrini, A. Ess, K. Schindler, and L. van Gool. 2009. You'll never walk alone: Modeling social behavior for multi-target tracking. In *2009 IEEE 12th International Conference on Computer Vision*. 261–268. https://doi.org/10.1109/ICCV.2009.5459260 ISSN: 2380-7504.

[26] Amir Rasouli, Iuliia Kotseruba, Toni Kunic, and John K. Tsotsos. 2019. PIE: A Large-Scale Dataset and Models for Pedestrian Intention Estimation and Trajectory Prediction. 6262–6271. https://openaccess.thecvf.com/content_ICCV_2019/html/Rasouli_PIE_A_Large-Scale_Dataset_and_Models_for_Pedestrian_Intention_Estimation_ICCV_2019_paper.html

[27] Simon Romero. 2022. Pedestrian Deaths Spike in U.S. as Reckless Driving Surges. *The New York Times* (Feb. 2022). https://www.nytimes.com/2022/02/14/us/pedestrian-deaths-pandemic.html

[28] Amir Sadeghian, Vineet Kosaraju, Ali Sadeghian, Noriaki Hirose, Hamid Rezatofighi, and Silvio Savarese. 2019. SoPhie: An Attentive GAN for Predicting Paths Compliant to Social and Physical Constraints. 1349–1358. https://openaccess.thecvf.com/content_CVPR_2019/html/Sadeghian_SoPhie_An_Attentive_GAN_for_Predicting_Paths_Compliant_to_Social_CVPR_2019_paper.html

[29] Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. 2020. Trajectron++: Dynamically-Feasible Trajectory Forecasting with Heterogeneous Data. In *Computer Vision – ECCV 2020 (Lecture Notes in Computer Science)*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer International Publishing, Cham, 683–700. https://doi.org/10.1007/978-3-030-58523-5_40

[30] Angie Schmitt. 2020. *Right of Way: Race, Class, and the Silent Epidemic of Pedestrian Deaths in America*. Island Press; Illustrated edition.

[31] Ramya Srinivasan and Ajay Chander. 2021. Biases in AI systems. *Commun. ACM* 64, 8 (Aug. 2021), 44–49. https://doi.org/10.1145/3464903

[32] Magdalena I. Tolea, Paul T. Costa, Antonio Terracciano, Michael Griswold, Eleanor M. Simonsick, Samer S. Najjar, Angelo Scuteri, Barbara Deiana, Marco Orrù, Marco Masala, Manuela Uda, David Schlessinger, and Luigi Ferrucci. 2010. Sex-Specific Correlates of Walking Speed in a Wide Age-Ranged Population. *J Gerontol B Psychol Sci Soc Sci* 65B, 2 (March 2010), 174–184. https://doi.org/10.1093/geronb/gbp130

[33] Sahil Verma and Julia Rubin. 2018. Fairness Definitions Explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*. 1–7. https://doi.org/10.1145/3194770.3194776

[34] Chuhua Wang, Yuchen Wang, Mingze Xu, and David J. Crandall. 2022. Stepwise Goal-Driven Networks for Trajectory Prediction. *IEEE Robotics and Automation Letters* 7, 2 (April 2022), 2716–2723. https://doi.org/10.1109/LRA.2022.3145090 Conference Name: IEEE Robotics and Automation Letters.

[35] Gang Wang, Shijia Pan, and Susu Xu. 2021. Decoupling the unfairness propagation chain in crowd sensing and learning systems for spatio-temporal urban monitoring. In *Proceedings of the 8th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*. 200–203.

[36] Heng Wang, Bin Wang, Bingbing Liu, Xiaoli Meng, and Guanghong Yang. 2017. Pedestrian recognition and tracking using 3D LiDAR for autonomous vehicle. *Robotics and Autonomous Systems* 88 (2017), 71–78.

[37] Benjamin Wilson, Judy Hoffman, and Jamie Morgenstern. 2019. Predictive Inequity in Object Detection. https://doi.org/10.48550/arXiv.1902.11097 arXiv:1902.11097 [cs, stat].

[38] Susu Xu, Xinlei Chen, Xidong Pi, Carlee Joe-Wong, Pei Zhang, and Hae Young Noh. 2019. ilocus: Incentivizing vehicle mobility to optimize sensing distribution in crowd sensing. *IEEE Transactions on Mobile Computing* 19, 8 (2019), 1831–1847.

[39] Susu Xu, Xinlei Chen, Xidong Pi, Carlee Joe-Wong, Pei Zhang, and Hae Young Noh. 2019. Vehicle dispatching for sensing coverage optimization in mobile crowdsensing systems. In *2019 18th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. IEEE, 311–312.

[40] Yu Yao, Ella Atkins, Matthew Johnson-Roberson, Ram Vasudevan, and Xiaoxiao Du. 2021. BiTraP: Bi-Directional Pedestrian Trajectory Prediction With Multi-Modal Goal Estimation. *IEEE Robotics and Automation Letters* 6, 2 (April 2021), 1463–1470. https://doi.org/10.1109/LRA.2021.3056339 Conference Name: IEEE Robotics and Automation Letters.

[41] Han Zhao. 2021. Costs and Benefits of Fair Regression. http://arxiv.org/abs/2106.08812 arXiv:2106.08812 [cs, stat].

[42] Motao Zhu, Songzhu Zhao, Jeffrey H. Coben, and Gordon S. Smith. 2013. Why more male pedestrians die in vehicle-pedestrian collisions than females: a decompositional analysis. *Inj Prev* 19, 4 (Aug. 2013), 227–231. https://doi.org/10.1136/injuryprev-2012-040594