Demonstrating Bias in Predictive Policing Tools Via Simulation Using GAN

Members: Kiran Prabhu, Samin Semsar, Uzma Hasan, Pronob Kumar Barman

Abstract

As crime rates continue to rise, the responsibility of ensuring community safety falls heavily on law enforcement officers. While increasing police presence at crime locations can help deter criminal activities to some extent, the accuracy of predictive policing algorithms in crime prediction remains a crucial question. These algorithms heavily rely on historical data to forecast crimes. However, empirical and theoretical research has highlighted a concerning feedback loop in predictive policing algorithms. The loop occurs when the model's estimates of crime probabilities influence the allocation of police resources, leading to a higher likelihood of detecting crimes in those specific areas. Consequently, this may trigger further resource allocation, resulting in an additional increase in the probability of discovering crimes. To address this issue, our research advocates for the utilization of Generative Adversarial Network (GAN) models to demonstrate the bias in existing predictive models by analyzing latitude and longitude data of crime locations from 911 calls made in 2019. Additionally, a Noisy-OR result is taken into account, introducing a level of randomness to the decision-making process. The results obtained through the GAN model provide valuable insights into the presence of bias within the detected crimes using the 2019 dataset. To further validate and strengthen our findings, we intend to extend our analysis by running the GAN model on multiple years' worth of data. By examining the consistency and patterns of bias across different time periods, we aim to gain a comprehensive understanding of the underlying biases in predictive policing models and contribute to the ongoing efforts towards achieving fairness and impartiality in resource allocation for law enforcement.

Introduction

In recent years, the field of predictive policing has emerged as a promising approach to combating crime and enhancing public safety. By leveraging advanced data analytics and machine learning techniques, predictive policing algorithms aim to identify crime patterns and allocate police resources effectively. However, the reliance on historical crime data in these algorithms raises concerns about potential biases and their impact on police allocation. Biases present in the data, such as underreporting or over-policing in certain communities, can lead to skewed predictions and perpetuate inequalities. To address these challenges, this research advocates the utilization of Generative Adversarial Network (GAN) models in the context of predictive policing. GANs have demonstrated their capability to detect and mitigate biases in various domains. By employing the GAN model, we aim to identify the biases that may arise from the historical crime data. GANs offer a unique approach by generating synthetic data that closely resembles the real crime data, allowing for a comprehensive analysis of biases and their impact on the predictions. This capability would enable law enforcement agencies to identify hotspots and allocate police resources more equitably, ensuring that the vulnerable communities are not disproportionately targeted or overlooked. Through the integration of the GAN model, predictive policing can become a powerful tool in enhancing public safety.

Problem Definition/Research Objective

Predictive models have emerged as a valuable tool for law enforcement agencies, providing them with the ability to automate and optimize the allocation of police resources to locations that are more likely to experience crimes. By leveraging historical data and machine learning techniques, these models aim to identify crime patterns and allocate resources accordingly, streamlining the manual process of patrolling. However, it is crucial to recognize that these predictive policing models are not immune to biases. Biases present in the historical data can result in unequal treatment and perpetuate disparities within the criminal justice system.

To address the issue, this study advocates for the utilization of the Generative Adversarial Network (GAN) as an advanced and innovative approach for detecting bias in current policing models. GANs, with their ability to generate synthetic data and analyze patterns, offer a powerful framework for identifying and understanding biases that may be embedded within predictive policing algorithms. By leveraging the power of GANs, we aim to detect the biases in these models.

Related Work

The literature on predictive policing and algorithmic bias is substantial, and it has received a lot of attention in recent years. Lum and Isaac's fundamental work (2016) identified a potential feedback loop in predictive policing algorithms that could lead to unwarranted over-policing in some places. Their findings emphasized the importance of corrective methods to counteract this inherent bias. Ensign et al. (2018) conducted more research into the underlying structural bias in predictive policing, suggesting that these algorithms may perpetuate and intensify historical patterns of discrimination.

GANs have lately found use in detecting biases in machine learning algorithms. GANs were first conceptualized by Goodfellow et al. (2014) and have since been used in a variety of disciplines, including detecting bias in data and techniques. For example, Caliskan et al. (2017) demonstrated gender and ethnic bias in word embeddings using a GAN-like approach, whereas Sun et al.

(2020) employed GANs to generate synthetic datasets for assessing bias in healthcare algorithms. Xu et al. (2019) designed the conditional tabular GAN (CTGAN) that can generate synthetic tabular data by addressing the challenges that tabular data might have. It uses a conditional generator and training-by-sampling to deal with the imbalanced discrete columns. It also applies the mode-specific normalization to overcome the problems of non-Gaussian and multimodal distributions. CTGANs can be used for data augmentation when the training data is too small for any model.

Jin et al. (2019) processed some regional crime dataset having data of fifteen years to learn latent representations using variational auto-encoder (VAEs). They proposed the Crime Generative Adversarial Network (Crime-GAN) which is formulated as a new crime forecasting model for four types of crime which integrates sequence to sequence structure and Wasserstein adversarial loss. Recently, GANs have been used for the task of debiasing biased classifiers. Mandis et al. (2021) used a GAN to debias a multi-class machine learning classifier that predicts a person's income based on their race and gender. Being inspired from classical GAN architecture, two neural networks are formed: a predictive network that takes in a person's features, excluding race and gender, to predict their income; and an adversarial network that infers the person's race and gender from the predictive network.

There is minimal yet promising research in the field of predictive policing. Carton et al. (2020) proposed a fairness-aware dispatch model that used machine learning to more evenly deploy police resources. They did not, however, use GANs or other bias detection tools. As a result, our study contributes to this emerging field by explicitly simulating and highlighting bias in predictive policing algorithms using GANs. This method contributes to a better understanding of the potential consequences of these biases and, more significantly, supports the development of strategies to mitigate them in future implementations.

Dataset

1. Neighborhood data was downloaded from here.
2. The neighborhood boundaries by downloading the KML file from here.
3. Crime data downloaded from here.

Summary of the crime dataset:

1. **CrimeDateTime** (mm/dd/yy format) The calls that came in through 911 after the incident
2. **CrimeCode** - Each of the crime codes represent the type of the crime (Ex: Burglary = 5D, Robbery = 3JF, Auto Theft = 7A, Shooting = 9S)
3. **Location** - Address of where the incident occurred.
4. **Description** - Type of the crime (one-on-one mapping with crime code, e.g. Burglary = 5D)

5. **ZipCode** - Zip code of the crime location
6. **Inside_Outside** - Whether the crime occurred inside (house, store etc) or outside (on the street, backyard etc)
7. **Weapon -** Not all the crimes in the dataset involves a weapon but some of them do, e.g. Firearm, Knife etc
8. **Post -** Indicates the smallest police geographic area - each post has a dedicated police post (Deployment of the police officers to locations for patrolling)
9. **District** - (administrative division - e.g. Eastern, Northern, Central etc)
10. **Neighborhood** - (name of the places)
11. **Premise** - Street, bus, store/outlet etc of where the incidents took place
12. **Latitude**
13. **Longitude**

The features used as input for the GAN model were::

1. Latitude
2. Longitude

Methodology

The primary objective is to identify hotspots that are at higher risk of experiencing crime. By analyzing latitude and longitude data of 2019 crime locations extracted from the crime dataset available on the Baltimore City website, this study aims to identify crime hotspots and deploy police officers accordingly.

The integration of the GAN model into the predictive policing system offers the potential to detect the crimes. GAN is a type of machine learning model that consists of two components: a generator and a discriminator. The generator's role is to create new data samples that resemble the real data. It learns to generate realistic samples by transforming random input (often called noise) into data that looks real. The discriminator on the other hand acts like a judge. It learns to distinguish between real data and the generated data. During the training process, both the generator and the discriminator play a game against each other. The generator tries to produce samples that can fool the discriminator into thinking they are real, while the discriminator aime to accurately classify real and generated samples. Through this adversarial process, both the generator and the discriminator improve overtime. The ultimate goal of the GAN model is to train the generator to produce high quality data that is indistinguishable from the real data.

The simulation runs 11 times. Once for each month, starting from the month of February. The locations of detected crimes from the previous month serve as the input to the GAN model, enabling the identification of hotspots for the current month's resource deployment. Moreover, by considering both the number of police officers in certain proximity to the crime location and a Noisy-OR result in decision-making, the determination of crime detection incorporates randomness. We set the crime detection radius to 700 feet, assuming that if there were a police officer patrolling within a 700 feet radius of a crime spot, there would be a high chance for crime detection. The availability of a larger number of police officers within the detection radius provides a higher probability for detection, and finally, using Noisy-OR the algorithm considers each crime to be either detected or undetected.

In the event that a crime remains undetected, we proceed to perform an additional round of Noisy-OR coin tossing, considering the probability of a crime being reported as 52.1% (as reported by the Pew Research Center, an American think tank organization). This step allows us to determine whether the crime is reported or not.
https://www.pewresearch.org/short-reads/2020/11/20/facts-about-crime-in-the-u-s/

By incorporating the detected crimes into the reported crime dataset, we augment the data for the subsequent month's GAN analysis, enabling us to generate the police locations for the following month. Multiple iterations of the year-long simulation are conducted, varying the inputs provided to the GAN model. In some instances, the GAN model is exclusively trained on the detected crimes from the previous month, while in other instances, it is trained on the reported crimes dataset. This approach allows for a comparative analysis of the results obtained from the different training inputs.

Results

We first reviewed the generated maps after running the simulation. As provided below, Fig. 1 shows the real crime locations in December 2019 in Baltimore City, plus the police locations that were chosen by the Gan to have the police officers dispatched there. The pink spots on the map refer to undetected crime locations in black neighborhoods, gray spots, undetected crime locations in white neighborhoods, light green spots, undetected crime locations in neighborhoods with neither black nor white as the majority of residents, dark red spots refer to detected crime locations in black neighborhoods, black spots, detected crime locations in white neighborhoods, and finally, dark green spots refer to detected crime locations in neighborhoods that have neither black nor white as the majority. The blue markers on the map are the police locations. Note that the dark colored spots, which are for the detected crimes, aren't very clear since big police blue markers are around them. Fig. 1 is the result of dispatching police forces based on detected crimes from the previous month for February 2019, while Fig. 2 is the same but for December of 2019. You see that the bias gets more prominent at the end of the year. Fig. 3 is the result of police distribution based on reported crimes from the previous month for December 2019. Comparing Fig. 3, it is clear that the bias is a bit less when we use reported crimes instead of detected ones. However, it doesn't remove the bias by any means.

As you can see, whether we use detected crime or reported crime in both cases, there's an obvious bias since all 60 police locations seem to have gathered in some neighborhoods while so many other neighborhoods are left with no forces.

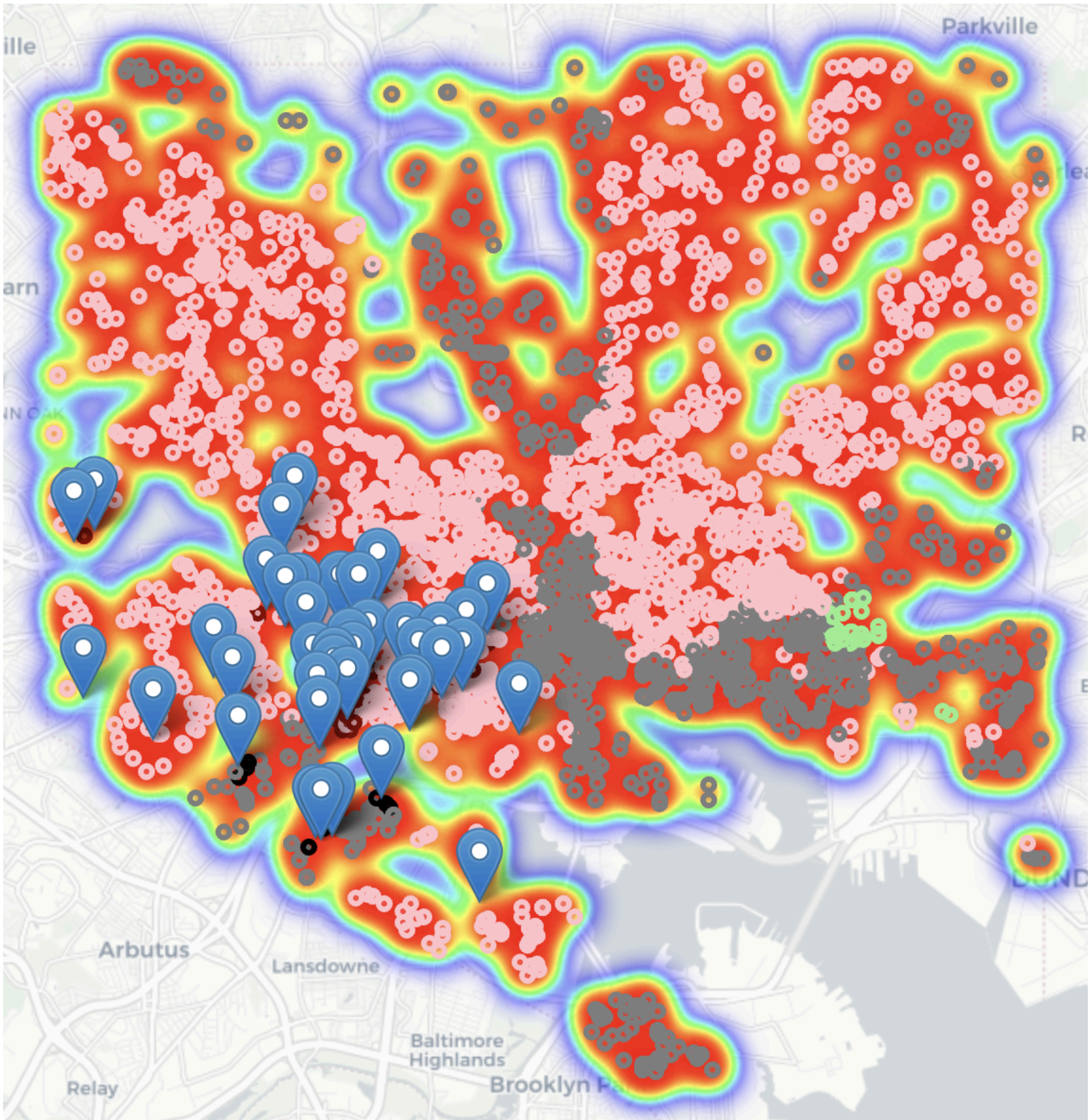


Fig.1 using detected crimes month 2

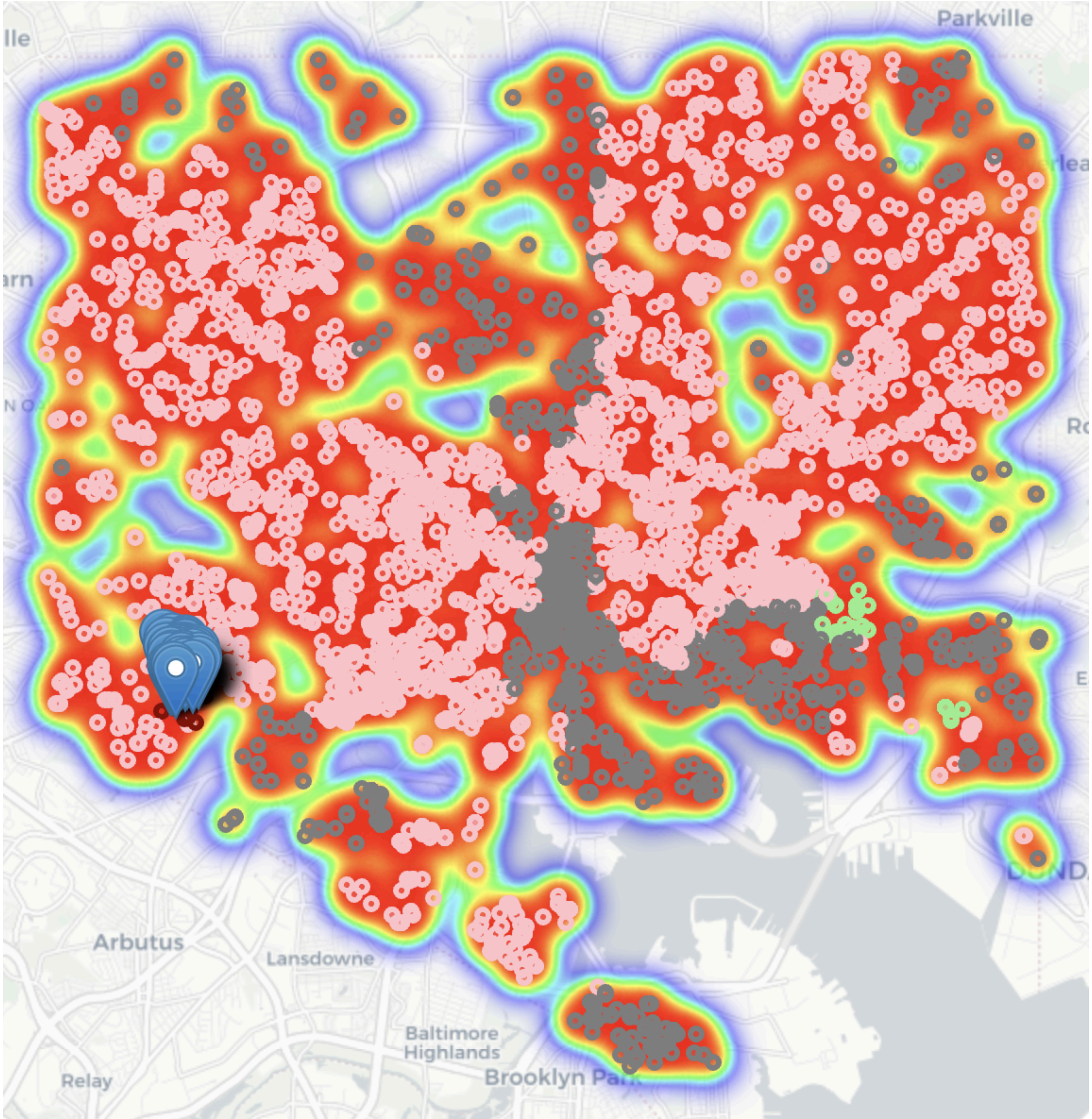


Fig 2. using detected crimes month 12

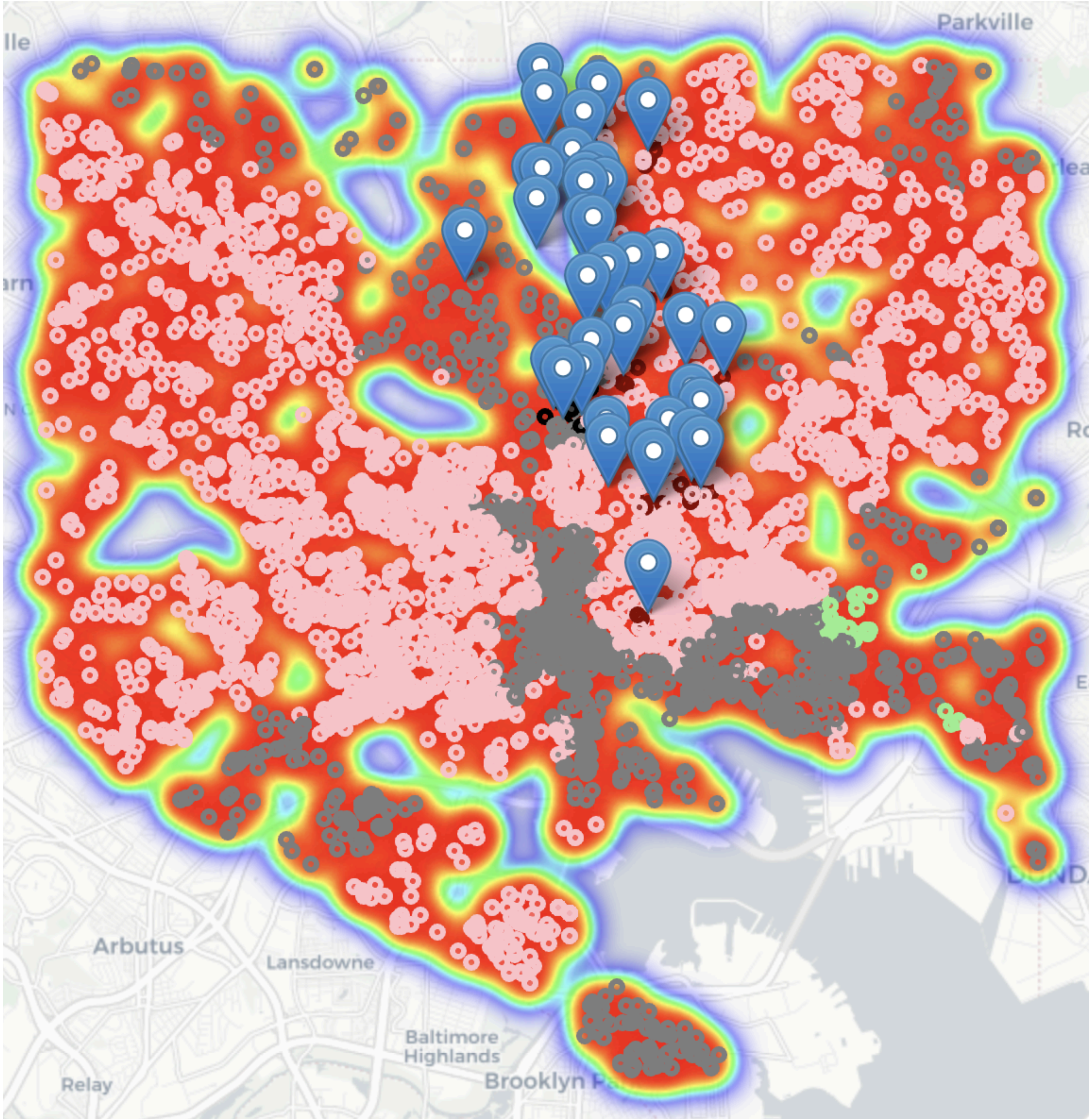
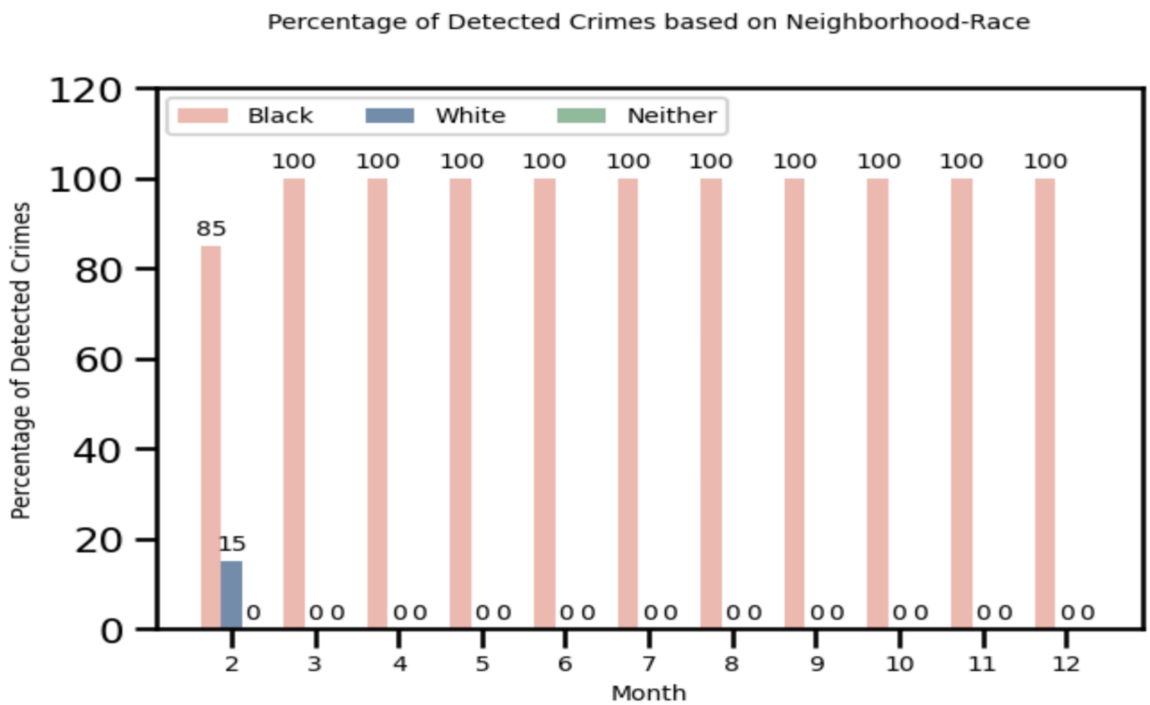
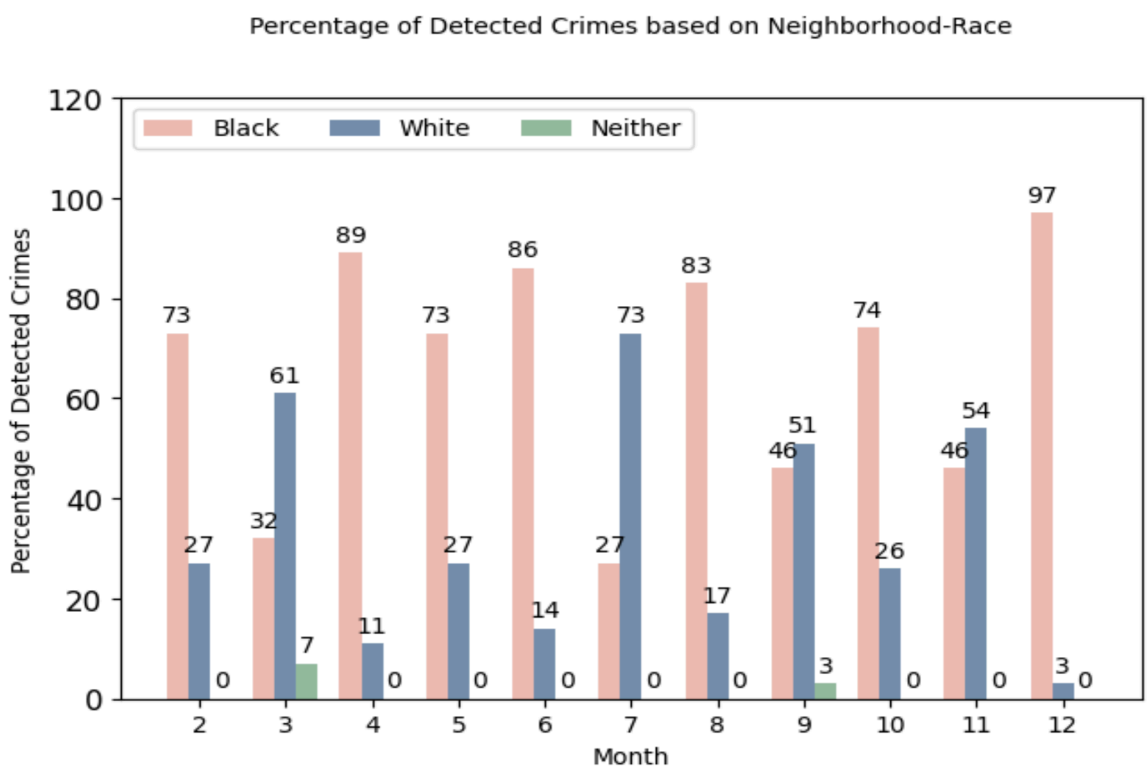


Fig 3. using reported crime month 12

In the bar charts below, one can see that using detected crimes, the bias is very much towards black neighborhoods, while in bar chart 2, the bias keeps shifting between white and black neighborhoods but is still mostly leaning towards black neighborhoods.



Barchart 1. Percentage of detected crimes through the months of the year 2019, using detected crimes



Barchart 2. Percentage of detected crimes through the months of the year 2019, using reported crimes

Conclusion and Future work

Previous studies have used Generative Adversarial Networks (GANs) for debiasing, and we hope to use this method to debias the police distribution by modifying our GAN structure in future work. When trained on a biased dataset, GANs can produce less biased samples. However, it is crucial to understand that GAN-based debiasing is not a universal solution. The quality and representativeness of the training data significantly impact the effectiveness of the debiasing approach. Additionally, biases in the dataset may not be easy to remove, and the GAN itself may add new biases to the generated samples. Debiasing the police distribution with GANs is a challenging and complex task. Ensuring that the training dataset is diverse and representative of the population is vital. Furthermore, it may be necessary to include domain knowledge and other methods to ensure that the generated samples are unbiased and meaningful. Knowing the challenges ahead helps with the planning and execution of a well-thought-out algorithm, ultimately increasing the likelihood of success. Therefore, our next course of action is to conduct a thorough review of the literature on the challenges associated with debiasing using GANs.

References

1. Moore, J., Pfeiffer, J., Wei, K., Iyer, R., Charles, D., Gilad-Bachrach, R., ... & Manavoglu, E. (2018). Modeling and Simultaneously Removing Bias via Adversarial Neural Networks. *arXiv preprint arXiv:1804.06909*.

2. Zhang, B. H., Lemoine, B., & Mitchell, M. (2018, December). Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 335-340).

3. Jin, G., Wang, Q., Zhao, X., Feng, Y., Cheng, Q., & Huang, J. (2019, December). Crime-gan: A context-based sequence generative network for crime forecasting with adversarial loss. In *2019 IEEE International Conference on Big Data (Big Data)*(pp. 1460-1469). IEEE.

4. Mandis, I. S. (2021). Reducing Racial and Gender Bias in Machine Learning and Natural Language Processing Tasks Using a GAN Approach. *International Journal of High School Research*, *3*(6), 17-24.

5. Louppe, G., Kagan, M., & Cranmer, K. Learning to pivot with adversarial networks (2016). arXiv preprint arXiv:1611.01046.

6. Gao, N., Xue, H., Shao, W., Zhao, S., Qin, K. K., Prabowo, A., ... & Salim, F. D. (2022). Generative adversarial networks for spatio-temporal data: A survey. ACM Transactions on Intelligent Systems and Technology (TIST), 13(2), 1-25.

7. Lum, K., & Isaac, W. (2016). To Predict and Serve? Significance, 13(5), 14-19. https://doi.org/10.1111/j.1740-9713.2016.00960.x

8. Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., & Venkatasubramanian, S. (2018). Runaway feedback loops in predictive policing. Proceedings of Machine Learning Research, 81, 1-12. https://arxiv.org/abs/1706.09847

9. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. Advances in neural information processing systems, 27, 2672-2680. https://papers.nips.cc/paper/5423-generative-adversarial-nets

10. Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. Science, 356(6334), 183-186. https://doi.org/10.1126/science.aal4230

11. Sun, T., Hoffman, J., & Grimmer, J. (2020). Simulating counterfactual representations of race for fairness in machine learning. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 314-324). https://doi.org/10.1145/3351095.3372854

12. Carton, S., Zeitzoff, T., & Anderson, A. (2020). Fair policing for the fair city: Mapping and assessing equity in the spatial allocation of police resources. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 304-313). https://doi.org/10.1145/3351095.3372873

13. Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling tabular data using conditional gan. Advances in Neural Information Processing Systems, 32.