AI and Ethics
https://doi.org/10.1007/s43681-024-00541-3

ORIGINAL RESEARCH

# Algorithmic fairness in predictive policing

**Ahmed S. Almasoud**[1] · **Jamiu Adekunle Idowu**[2]

Received: 31 January 2024 / Accepted: 29 July 2024


**Abstract**
The increasing use of algorithms in predictive policing has raised concerns regarding the potential amplification of societal biases. This study adopts a two-phase approach, encompassing a systematic review and the mitigation of age-related biases in predictive policing. Our systematic review identifies a variety of fairness strategies in existing literature, such as domain knowledge, likelihood function penalties, counterfactual reasoning, and demographic segmentation, with a primary focus on racial biases. However, this review also highlights significant gaps in addressing biases related to other protected attributes, including age, gender, and socio-economic status. Additionally, it is observed that police actions are a major contributor to model discrimination in predictive policing. To address these gaps, our empirical study focuses on mitigating age-related biases within the Chicago Police Department's Strategic Subject List (SSL) dataset used in predicting the risk of being involved in a shooting incident, either as a victim or an offender. We introduce Conditional Score Recalibration (CSR), a novel bias mitigation technique, alongside the established Class Balancing method. CSR involves reassessing and adjusting risk scores for individuals initially assigned moderately high-risk scores, categorizing them as low risk if they meet three criteria: no prior arrests for violent offenses, no previous arrests for narcotic offenses, and no involvement in shooting incidents. Our fairness assessment, utilizing metrics like Equality of Opportunity Difference, Average Odds Difference, and Demographic Parity, demonstrates that this approach significantly improves model fairness without sacrificing accuracy.

**Keywords** Predictive policing · Algorithmic fairness · Mitigating age bias · Conditional score recalibration

## 1 Introduction

Over the past few years, the adoption of Artificial Intelligence and Machine Learning technologies in the security sector has seen a significant rise. A crucial application within this area is predictive policing systems, which are designed to predict potential criminal activities using data-driven methodologies [23]. Aside from predicting where a crime might take place, these algorithms can also identify individuals who could potentially be involved in criminal activities in the future. Furthermore, with the advancement of computer vision in the security sector, AI applications for real-time crime detection and prevention are increasing [1, 29, 32]. Alongside predictive policing, recidivism algorithms have also gained traction. These algorithms focus on predicting the likelihood of a previously convicted individual reoffending. Just like predictive policing, these algorithms hold the promise of optimizing the justice system but come with their own set of challenges especially bias which needs to be addressed for AI to be a true enabler of sustainable development [14, 29].

The quality and representativeness of the data that informs both predictive policing and recidivism algorithms have become a point of contention. There are concerns about data that may be skewed thereby leading algorithms to perpetuate or even exacerbate these biases in their predictions [6, 23]. This is evident in datasets used for crime detection, which might not adequately represent all societal groups [26]. Also, there have been instances in history when risk assessments were influenced by sensitive features like race, nationality, and skin color [4]. Beyond this, other factors like

✉ Jamiu Adekunle Idowu
ucabaid@ucl.ac.uk

Ahmed S. Almasoud
almasoud@psu.edu.sa

[1] Department of Information System, Prince Sultan University, Riyadh, Saudi Arabia

[2] Centre for Artificial Intelligence, Department of Computer Science, University College London, London, United Kingdom

Published online: 02 September 2024

socio-economic status and age can also introduce unwanted racial or demographic biases [28].

The implications of such biased algorithms are far-reaching. They can result in marginalized communities facing further disadvantages, or individuals suffering unwarranted repercussions like false arrests based on AI's misclassifications [2, 9]. Recognizing these challenges, researchers have proposed various mitigation strategies and fairness metrics. Therefore, in our systematic review, we discussed the methodologies proposed to enhance algorithmic fairness in both predictive policing and recidivism. The primary objectives are as follows:

1. To analyze the fairness metrics, sensitive features and fairness strategies employed in existing literature towards mitigating bias in policing algorithms.
2. To mitigate age-related biases in predictive policing algorithms.

## 2 Objective 1: systematic literature review

In this section, we address our first research objective by conducting a systematic literature review (SLR) to analyze fairness metrics, sensitive features and fairness strategies employed in existing literature. Our systematic literature review includes papers that meet a set of predetermined eligibility criteria.

### 2.1 Methods for SLR

We used the PRISMA guideline, updated in 2020 for the review process [21]. The stages involved in this SLR are outlined below:

#### 2.1.1 Stage 1: data sources and search strategy

For the primary source of our systematic search, we chose Scopus due to its vast collection of over 70 million publications, coupled with its high recall, precision, and reproducibility [34]. In addition, IEEE Xplore and ScienceDirect are used as supplementary search systems [34]. The keywords used for the search strategy are presented in Table 1.

#### 2.1.2 Stage 2: data sources and search strategy

This section sets out the criteria we applied when selecting studies for the systematic literature review. We screened studies in a three-step process: initially examining titles and keywords, followed by abstracts, and then a thorough reading of the full text. Studies fitting our criteria proceeded to Stage 3 for further evaluation.

#### 2.1.3 Inclusion criteria

- Articles that directly investigate bias issues in policing or recidivism algorithms.
- Publications from 2015 or later.
- Research materials including peer-reviewed journals, conference papers, and white papers that have defined research questions.

#### 2.1.4 Exclusion criteria

- Non-research materials such as newsletters, magazines, posters, invited talks, panels, keynotes, and tutorials.
- Duplicate research paper or article.

#### 2.1.5 Stage 3: quality assessment

In our review, we used a specific set of questions to evaluate the quality and relevance of each study. A total of five key questions were employed to assess quality. Studies meeting at least 3 out of these 5 questions were included for deeper analysis.

QA1. Did the study have clearly defined research questions and methodology?
QA2. Did the study develop an algorithm specifically for crime prediction or recidivism?
QA3. Did the study incorporate considerations of fairness during the algorithm development process, ensuring non-discrimination based on factors like gender, race, age, etc.?
QA4. Was the study based on a real-world dataset?

**Table 1** Search Keywords

| Term | Keywords |
|---|---|
| Fairness | fair* OR bias OR trust* OR responsible OR equity OR "social justice" OR discrimination AND |
| Police/recidivism | police OR "law enforcement" OR policing OR "public safety" OR crim* OR recidivism AND |
| Algorithm | algorithm* OR "machine learning" OR models OR "artificial intelligence" OR predictive OR automated |

QA5. Did the study report the performance of the algorithm using suitable metrics such as accuracy, F1 score, AUC-ROC, precision, recall, etc.?

## 2.2 Results

From the initial search, we identified 587 relevant studies. Out of these, 293 were sourced from Scopus, 185 from SpringerLink, and 109 from IEEE Xplore. Using EndNote, a reference management tool, we identified and removed 89 duplicate studies. This left us with 498 studies ready for title and abstract screening. Upon examining the titles and keywords, we excluded 426 studies. This led us to thoroughly read the abstracts of the remaining 72 articles. From this group, 34 were found not to meet our criteria, leaving 38 for full-text examination. After this detailed review, 20 more articles were set aside, narrowing the list down to 18 as shown in Table 2.

Following our quality assessment, only 15 of these studies were deemed suitable for our review. Figure 1 provides a visual representation of this selection process using a PRISMA flow chart.

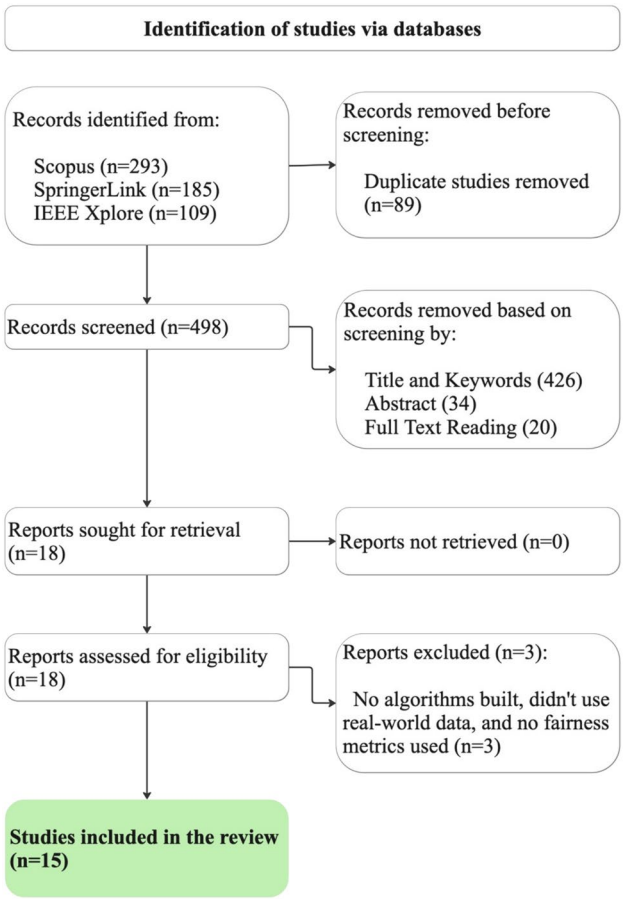


**Fig. 1** Flow diagram of study selection using PRISMA

## 2.3 Discussion

### 2.3.1 Fairness metrics

This review uncovers a variety of methodologies and metrics employed in the literature for measuring fairness of policing and recidivism algorithms, showcasing the complexity

**Table 2** Papers selected for Quality Assessment Evaluation

| Paper | Reference | Title |
|---|---|---|
| S1 | [8] | Evaluating Fairness in Predictive Policing Using Domain Knowledge |
| S2 | [5] | Improving fairness in criminal justice algorithmic risk assessments using optimal transport and conformal prediction sets |
| S3 | [13] | Predictive policing and algorithmic fairness |
| S4 | [22] | Data augmentation for fairness-aware machine learning: Preventing algorithmic bias in law enforcement systems |
| S5 | [27] | Is the data fair? An assessment of the data quality of algorithmic policing |
| S6 | [20] | Cohort bias in predictive risk assessments of future criminal justice system involvement |
| S7 | [19] | A penalized likelihood method for balancing accuracy and fairness in predictive policing |
| S8 | [25] | AI For Bias Detection: Investigating the Existence of Racial Bias in Police Killings |
| S9 | [30] | Accuracy and fairness in a conditional generative adversarial model of crime prediction |
| S10 | [12] | Artificial fairness? Trust in algorithmic police decision-making |
| S11 | [31] | Crowdsourcing perceptions of fair predictors for machine learning: A recidivism case study |
| S12 | [16] | Singular race models: addressing bias and accuracy in predicting prisoner recidivism |
| S13 | [17] | Algorithmic bias in recidivism prediction: A causal perspective |
| S14 | [15] | Accuracy, Fairness, and Interpretability of Machine Learning Criminal Recidivism Models |
| S15 | [24] | Case study: predictive fairness to reduce misdemeanor recidivism through social service interventions |
| S16 | [3] | A review of predictive policing from the perspective of fairness |
| S17 | [7] | Fairness, accountability and transparency: notes on algorithmic decision-making in criminal justice |
| S18 | [33] | Achieving equity with predictive policing algorithms: a social safety net perspective |

and diversity of the field. For example, S1 employed five distinct metrics; S2 used predictive parity; S9 implemented the calibration test; S12 focused on False Positive Rates and False Negative Rates, and S14 utilized fairness metrics from the Aequitas toolkit. The five fairness metrics used by S1 are briefly explained below:

- *Statistical parity difference (SPD):* Measures the *difference* in favorable outcomes between the unprivileged and privileged groups. If SPD is zero, it indicates perfect fairness, but a non-zero value indicates some level of disparity between the groups [8].
- *Disparate impact (DI):* Measures the *ratio* of favorable outcomes for the unprivileged group to the favorable outcomes for the privileged group. A value of 1 indicates perfect fairness (i.e., both groups receive favorable outcomes at the same rate). A value less than 1 suggests that the unprivileged group is less likely to receive a favorable outcome compared to the privileged group, while a value greater than 1 indicates the opposite [8].
- *Average odds difference:* Measures the average difference of false and true positive rate between the unprivileged and privileged group [8].
- *Equal opportunity difference:* Measures the difference of true positive rates between the privileged and unprivileged groups [8].
- *Theil index:* A measure of the inequality in benefit allocation for individuals.

Meanwhile, S2 adopted prediction parity as it is good for capturing group fairness [5]. On the other hand, S9 employed a calibration test to measure algorithmic fairness, with residential income as the protected attribute. The fairness measure involved assessing the calibration over a data distribution, $D$. As an example, if 100 data points from $D$ predicted a crime occurrence with a 0.7 confidence level, then the model's calibration is validated if 70 of its predictions are accurate. This calibration assessment was executed across ten bins of equal prediction accuracy [30].

Another eligible study, S12 utilized False Positive Rate (FPR) and False Negative Rate (FNR) as metrics to identify biases within their models. A high FPR implies a significant number of non-recidivists being wrongly classified as potential reoffenders, leading to their undue detention. Conversely, a high FNR indicates a considerable number of actual recidivists incorrectly deemed fit for release, potentially resulting in preventable crimes. The study also explored the implications of these rates for different subpopulations, highlighting that biases manifest differently among groups. For instance, they observed that African American group members often exhibited characteristics of a higher false positive rate and a lower false negative rate compared to other populations [16].

In their study, S14 leveraged several metrics from the Aequitas toolkit. These metrics include Predicted Positive Rate Disparity (PPRD), which measures whether the numbers of positive predictions are on par across groups. Also, False Discovery Rate Disparity (FDRD) assessed the proportion of false positives relative to all predicted positives, while False Positive Rate Disparity (FPRD) focused on the ratio of false positives to the actual negatives across groups. Additionally, the study considered False Omission Rate Disparity (FORD), evaluating the false negatives relative to predicted negatives, and False Negative Rate Disparity (FNRD), which assessed the ratio of false negatives compared to the actual positives [15].

#### 2.3.2 Datasets and sensitive features

In Table 3, we make a comparison between the eligible papers based on the datasets and sensitive features they used. Most of the studies on Policing or Recidivism algorithmic fairness are based on datasets from the United States. They include Chicago's Strategic List data (S1), New York Police Department Crime Complaint data (S5), Indianapolis Crime Incident data (S7), Police Killings data (S8), Criminal Defendants of Broward County, Florida (S11), NIJ data from Georgia, and Los Angeles' Case Management data (S15). The only exception is S9 which used crime incidents dataset from SIEDCO, Bogotá, Colombia.

In terms of protected attribute being investigated, studies gave major attention to race as 12 of the 15 studies compared in the table investigated racial bias with the only exemption being S6, S9, and S10. Meanwhile, only two of the 15 studies investigated bias related to socio-economic status or income (S9 and S11). Similarly, only four out of 15 papers each examined bias related to gender (S1, S5, S11, and S14) and three for age (S1, S5 and S6). In addition, only four of the studies examined bias in more than one feature (S1, S5, S11, S14).

#### 2.3.3 *Bias* analysis and mitigation strategies

The analysis of the eligible papers shows that bias analysis methods and mitigation strategies commonly employed in policing and recidivism algorithms revolve around themes like data-centric strategies (S1, S4, S5, & S8), counterfactual reasoning (S2, S3, & S13), demographic segmentation (S6, S12, & S15), and adding penalty to likelihood function (S7). Table 4 presents some of the bias analysis techniques and mitigation strategies.

**2.3.3.1 Data-centric strategies (S1, S4, S5, and S8)** Several of the approaches directly target the dataset itself, reflecting the maxim that "better data leads to better predictions."

**Table 3** Comparison of selected papers based on sensitive features

| Paper | Datasets | Sensitive features | | | |
|---|---|---|---|---|---|
| | | Sex | Race | Status* | Age |
| S1 | Strategic Subject List (SSL) dataset from Chicago Police Department listing arrest of 398,684 people from August 1, 2012, to July 31, 2016 | ✓ | ✓ | | ✓ |
| S2 | A random sample of 300,000 offenders in the United States | | ✓ | | |
| S3 | Strategic Subject List (SSL) dataset from Chicago Police | | ✓ | | |
| S4 | RWF-2000, which contains two thousand video clips lasting five seconds each, half of them depicting violent acts and half of them normal activities | | ✓ | | |
| S5 | NYPD's Crime Complaint Data (January–June 2018) published on NYC.gov | ✓ | ✓ | | ✓ |
| S6 | Data from Project on human development in Chicago covering 1,057 individuals from four different age cohorts | | | | ✓ |
| S7 | Crime incident data from the city of Indianapolis, Indiana for the years 2012 and 2013. The four crime types include aggravated assault, robbery, motor vehicle theft, and burglary | | ✓ | | |
| S8 | 2013–2020 Police Killings" sheet from the Mapping Police Violence database | | ✓ | | |
| S9 | Dataset from SIEDCO that records all crime incidents in the city of Bogotá covering period 2016–2019. The data set has 183,813 reports divided into thefts, injures, and homicides | | | ✓ | |
| S10 | Online experiment involving 642 UK residents recruited from an online crowdsourcing platform | | | | |
| S11 | Public dataset containing information on criminal defendants of Broward County, Florida | ✓ | ✓ | ✓ | |
| S12 | Dataset assembled using the information and resources acquired from Florida Department of Corrections (FDOC) and the Florida Department of Law Enforcement (FDLE) | | ✓ | | |
| S13 | Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) data offer 2 years of data (2013–2014) | | ✓ | | |
| S14 | NIJ contains information about 25,835 individuals from the state of Georgia who were granted parole from 2013–2015 | ✓ | ✓ | | |
| S15 | Data extracts from the Los Angeles Attorney's case management system were provided for the project | | ✓ | | |

- *Incorporation of relevant external data:* S1 integrates domain knowledge into a predictive policing dataset by adding average rates of education, employment, and poverty from the 2014 census, aiming to minimize bias and fairness concerns. It found that integrating this domain knowledge has minimal impact on classification metrics, but notably improves fairness for all protected classes, particularly in the Theil index for race [8].
- *Data augmentation:* To counteract bias in law enforcement algorithms, S4 leverages a data augmentation strategy for the RWF-2000 dataset, which showed an overrepresentation of dark-skinned males in violent situations. Through techniques like Mask-RCNN and HRNet-w48, the study synthetically oversamples undersampled entities to achieve balance, particularly focusing on race. The results indicate that models trained on balanced datasets generalized better, with increased accuracy and reduced racial bias [22]. Similarly, S8 assigned class weights and oversampled minority class to address data imbalance.
- *Quality of the data:* According to S5, data quality can be as influential as the algorithm itself in determining outcomes as missing or incomplete data can lead to inaccurate or biased predictions. The paper's analysis of NYPD's Crime Complaint Data shows the data has a high imbalance, which can affect the reliability of predictions. Depending on the percentage of missing values, it might be advisable to either conduct a complete case analysis or discard certain variables from the dataset [27]. For instance, according to the recommendations in [18], if a dataset has less than 5% of its values missing, it might be appropriate to conduct a complete case analysis. This is because any resulting bias is likely, though not guaranteed, to be minimal. On the other hand, if over 40% of the values for a particular variable are missing, one might consider removing that variable from the analysis altogether.

**2.3.3.2 Counterfactual reasoning and causal analysis (S2, S3 and S13)** S2 improved fairness in risk assessment algorithms by employing counterfactual reasoning. First, it trains a stochastic gradient boosting algorithm on a more privileged protected class (White offenders). Then, it transports the joint predictor distribution from the less privileged class (Black offenders) to the more privileged class with the aim of treating the disadvantaged class members as if they are members of an advantaged class [5]. By using optimal transport, S3 removed the bias caused by differences at arrangement between the joint predictor distributions for Black and White offenders [13].

**Table 4** Bias Analysis and Mitigation Strategies

| Paper | Strategy/Analysis | Implementation | Result |
|---|---|---|---|
| S1 | Incorporating domain knowledge | Added features from external data (2014 census data)—average education rate, employment rate, poverty rate | Improved Theil index fairness for race while having minimal effect on accuracy |
| S2 | Counterfactual Reasoning for Improved Fairness | Model trained on privileged class; Transfer of joint predictor probability using optimal transport | Significant improvements in fairness for protected classes |
| S3 | Causal Analysis for Model Discrimination | Analysis of the causal relationships among various variables impacting PPAs, particularly focusing on the role of police actions | Police actions identified as the primary driver of model discrimination in predictive policing |
| S4 | Fairness-Aware Data Augmentation | Identify biases; Oversample underrepresented entities in video sequences | Improved accuracy; Reduced biases |
| S5 | Data Quality Analysis | Missingness analysis on NYPD crime complaint data | Identified high missingness; highlighted importance of data quality for fairness in algorithmic policing |
| S6 | Cohort Bias Analysis | Analyzed criminal histories across birth cohorts to assess predictive risk assessments | Revealed overprediction of arrest risk for younger cohorts; found bias across all racial group |
| S7 | Penalized Likelihood Approach for Demographic Parity | Added a penalty term to the likelihood function for police patrols | Introduced fairness aligning patrol levels with demographics; Accuracy decreased when prioritizing fairness |
| S8 | Class weights and data oversampling | Applied multiple ML models to an expansive police killing dataset; used class weights and overspampling to correct data imbalance | Achieved over 80% accuracy in classifying race of fatalities |
| S9 | Conditional GANs Architecture for Crime Prediction | ConvLSTM layers conditioned on past crime intensity maps, weekdays, holidays | Improved spatiotemporal predictions; Reduced bias with minimal accuracy loss |
| S10 | Online experiment, text-based vignettes, random allocation to scenarios | Tested public trust in police decisions made by human vs. algorithm, successful vs. unsuccessful outcomes, and individual vs. area-based scenarios | People viewed decisions made by algorithms as less fair and appropriate; successful algorithmic outcomes increased support for general police use of algorithms; trust influenced acceptability of algorithmic policing |
| S11 | Crowdsourcing perceptions of fair predictors | Recruited 90 crowdworkers to judge the inclusion of various predictors for recidivism; used a bot to facilitate discussion and voting | Diverse groups tended to agree more with the majority vote; visual evidence was crucial in decision-making |
| S12 | Singular Race Models | Segment dataset based on race and Train race-specific models with artificial neural networks | For most race-crime categories, improved accuracy over base models; Magnitude of bias increased |
| S13 | Causal reformulation of algorithmic fairness problem | Used the Neyman-Rubin potential outcomes framework to estimate algorithmic fairness from COMPAS data | Strong evidence of racial bias against African American defendants |
| S14 | Comparative analysis of ML models | Created and evaluated various ML models (decision tree, logistic regression, boosting classifiers, random forest, SVM, neural network) | Found trade-offs between accuracy, fairness, and interpretability; XGBoost had highest accuracy; Decision tree was most interpretable and fairest for gender; Random forest was fairest for both gender and race |
| S15 | Post-hoc Bias Detection and Mitigation | Adjusted score thresholds to balance recall across racial/ethnic groups in predictive models | Achieved better predictive fairness; identified trade-offs between equity and efficiency |

Also, a causal analysis by S3 on predictive policing algorithms (PPAs) found that police actions are the primary factor contributing to model discrimination. The research uncovered a vicious cycle: increased police deployment leads to higher arrest rates, which then increases reported crime rates, further justifying more police deployment [13]. The study considered a case study of the Chicago Police Department with variables A (arrest records), O (PPA output), I (PPA input and training), D (police actions), C (reported crime rate), and relationships shown in Fig. 2. A vicious cycle emerges between D, A, and C (Fig. 2), when PPAs are not involved. When PPAs are integrated, complex interactions among variables arise. Even if PPAs are banned, effectively eliminating I and O, as recommended by [11], the inherent bias in D persists. Therefore, D (police actions) is the major driver of model discrimination, confirming [10]'s findings that the racial disparity in Los Angeles Predictive Policing Algorithms "lies in policing, not the algorithm."

In addition, S13 analyzed the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) data using the Neyman-Rubin potential outcomes framework for causal inference. The analysis showed that the COMPAS data has racial bias against African American defendants.

**2.3.3.3 Demographic segmentation (S6, S12, and S15)** S12 developed Singular Race Models, a novel approach to reduce racially inspired bias in recidivism prediction, by segmenting the dataset based on race and training single race-based models. This approach increased prediction accuracy and analyzed race-related discrimination. However, it was noted that bias still existed in the two race-based cohorts even in the absence of race information, implying that many features in the dataset are correlated with race [16]. **Also,** S15's approach of developing individually tailored interventions based on demography further emphasizes the importance of understanding and catering to specific demographic groups [24]. In addition, S6 highlighted an essential challenge known as cohort bias. By analyzing criminal histories across different age cohorts, the study found that models trained on older cohorts tend to over-predict the likelihood of arrest for younger cohorts [20].

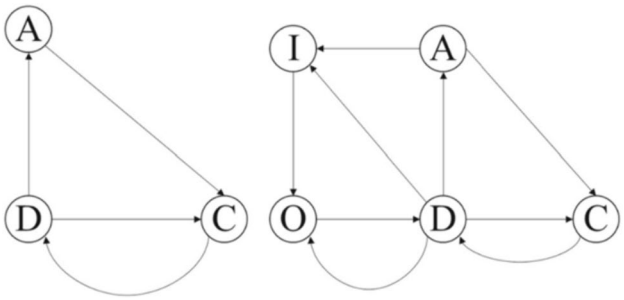


**Fig. 2** Casual relationships before (left) and after (right) PPA integration (S3)

**2.3.3.4 Fairness-accuracy trade-off (S4, S7, and S9)** A recurrent trend in these studies is the exploration of the balance between fairness and accuracy. S7 found a reduction in algorithmic accuracy when demographic fairness was achieved [19]. S9 introduced a fairness-accuracy balancing technique, which quantified the tradeoffs between accuracy and fairness of GANs model, successfully reducing bias with a marginal impact on accuracy [30]. Similarly, the data augmentation strategy by S4 increased accuracy and reduced racial bias.

**2.3.3.5 Others (S7, S9, 10, 11, and 14)** S7 introduced a penalized likelihood approach to achieve demographic parity in crime models. The proposed strategy adjusts the likelihood function with a penalty term, ensuring police patrol distribution among demographic groups is proportional to their population representation. When applied to crime data in Indianapolis, the model successfully implemented fairness in patrol distribution, but this came with a minor reduction in prediction accuracy [19]. Meanwhile, another study, S9 used a conditional GANs architecture to predict crimes in Bogotá, Colombia, and identified potential bias affecting vulnerable populations. To address fairness, S9 performed a calibration test conditional on income level (high income vs low income) as a protected variable [30]. Meanwhile, S10 examined public trust in algorithms versus human decision-making in policing, finding that algorithmic decisions are viewed as less fair and appropriate, yet successful use cases can enhance overall acceptance. S11 focused on crowdsourcing perceptions of algorithmic fairness, with participants evaluating whether predictors are fair or not. S14 conducted a comparative analysis of ML models for criminal recidivism, highlighting trade-offs between accuracy, fairness, and interpretability, with XGBoost showing the highest accuracy and decision trees being the most interpretable and fairest in gender bias.

### 2.3.4 Human perceptions

One way to approach algorithmic bias in policing is to include humans in the decision-making loop to bring a layer of understanding, validation, and accountability to the process. **S10** investigated public perception and trust levels towards algorithmic decision-making in policing, especially when compared with decisions made by human officers. The study conducted an online experiment involving 642 UK residents recruited from an online crowdsourcing platform. S10 found that people generally view decisions made by algorithms as less fair and appropriate compared to those made by police officers. However, this perception changes

with context. When exposed to instances of successful algorithmic decision-making, trust in the algorithm and its potential benefits increases. It is worth noting that the sample for this study was predominantly composed of younger and White-British individuals, which could have potential implications for the generalizability of these findings [12].

In a similar vein, S11 recruited 90 crowdworkers to identify predictors deemed 'fair' for algorithmic models. Participants, in both individual and group settings, judged the fairness of predictors for predicting recidivism. The study's design, which utilized a popular workplace collaboration tool, Slack, allowed for a structured yet dynamic exploration of participant's opinions. The results highlighted that participants in a more structured group setting (with an appointed leader) showed higher agreement with majority votes when assessing predictor fairness. Additionally, there were some predictors, like the defendant's race, that were identified for exclusion in both participants didn't seem to influence their voting behavior on demographic predictors, suggesting that individual biases may not have played a significant role in the decision-making process [31].

### 2.4 Recommendations based on SLR

In our systematic literature review, we have investigated fairness in policing and recidivism algorithms. The analysis of the selected papers revealed a range of strategies aimed at addressing inherent biases in these algorithms, highlighting the multidimensional nature of fairness in the domain. Based on our analysis, we make the following recommendations for researchers to consider in future:

1. *Human validation in algorithmic processes:* Considering findings from S10 and S11, it is recommended to incorporate human judgment in algorithmic processes. This ensures decisions benefit from both human intuition and machine efficiency. Additionally, structured group settings, as observed in S11, can be adopted for consistent fairness evaluations.
2. *Expand datasets beyond the US:* Given the predominance of U.S-based datasets as observed in studies like S1, S2, S3, S5, S6, S7, S11, S12, S14, and S15, researchers are encouraged to diversify geographical sources to enhance global relevance as the societal norms and legal structures influencing criminal behavior vary widely across countries [37]. Also, the U.S. context often differs from other regions in terms of legal systems, socio-economic conditions, and demographic compositions. Therefore, relying solely on U.S-based datasets can limit the generalizability of findings and fairness solutions.
3. *Expand the scope of protected attributes in bias investigations:* The emphasis on racial bias in the majority of the studies, as observed from 12 out of 13 papers referenced, highlights a narrow focus in bias investigations. In contrast, other critical attributes like socio-economic status or income, gender, and age received significantly less attention, with only two studies examining bias related to socio-economic status (S9 & S11) and only three studies each for gender (S1, S5 & S11) and age (S1, S5 & S6). Therefore, to ensure fairness across multiple dimensions with better generalizability of algorithmic solutions, it's essential for future research to pay attention to gender, socio-economic status, and age in addition to race.
4. *Tailor bias mitigation based on context:* Given the diverse strategies highlighted in S1, S2, S4, S7, S9, S12, and S18, it's advisable for researchers to critically analyze biases based on the specific context and application of policing and recidivism algorithms. Bias mitigation should also be tailored to fit the unique context, ensuring more relevant and effective fairness outcomes.

## 3 Objective 2: mitigating age-related *bias*

In this section, we implement the second objective of our paper which involves investigating and mitigating bias related to age in policing algorithms. The policing algorithm developed in this study is for predicting the risk of involvement in a shooting incident either as a victim or an offender and it is based on the Chicago Police Department's Strategic Subject List.

As highlighted in our SLR, age, socio-economic status, and gender are sensitive features often overlooked by previous studies on predictive policing. For instance, only 3 out of the 15 papers in our SLR examined bias related to age, 4 examined gender bias, and only 2 investigated bias related to socio-economic status compared to 12 out of 15 for race. While any of these understudied categories could be chosen towards addressing the gap identified in our SLR, the choice of focusing on age-related bias in this study is driven by the characteristics of the dataset being analyzed. Specifically, the Chicago Police Department's Strategic Subject List contains age and gender but lacks socio-economic status data. Additionally, the dataset shows significant bias with respect to age as reported in [8] and [37]. Therefore, we are focusing on mitigating age-related bias in this study. However, this does not negate the importance of socio-economic status or gender; rather, it addresses one part of the gaps while recognizing that others remain and should be explored in future research.

### 3.1 Dataset

This research used the Chicago Police Department's Strategic Subject List dataset which covers 398,684 people from

August 2012 to July 2016. The Strategic Subject List (SSL) score is a risk assessment score that reflects an individual's probability of being involved in a shooting incident either as a victim or an offender. The scores range from 0 (extremely low risk) to 500 (extremely high risk). According to [36], scores above 250 are deemed to be high risk. According to the Chicago Police Department, the SSL is intended to "rank individuals according to their probability of being involved in a shooting incident, either as an offender or a victim," with scores recalculated daily.

Meawhile, there is a huge discrepancy in the dataset that points to age bias as analysis revealed that age accounts for about 89% of the variance in SSL scores [36]. Specifically, 127,513 individuals have never been arrested or shot, but around 90,000 of them are deemed to be at high risk.

#### 3.1.1 Data pre-processing

Initially, the dataset has 398,684 data instances and 48 features. We applied preliminary pre-processing steps which involved dropping columns with significant missing values, dropping rows with missing values for longitude and latitude columns, among others. Afterwards, we followed up with additional pre-processing steps highlighted below:

1. The feature 'AGE AT LATEST ARREST' originally contained age ranges. We simplified this by categorizing individuals into two groups: those under 30 years old ('less than 20' and '20–30') were encoded as 1 and those 30 years or older were encoded as 0. This binary classification is based on our initial findings that the most pronounced differences in SSL scores were between these two age groupings. One critical consideration is that the split needs to be relatively balanced. The data points for each age range before the binary split were as follows:
   - Less than 20 yrs: 41,415 individuals
   - 20–30 yrs: 76,731 individuals
   - 30–40 yrs: 46,915 individuals
   - 40–50 yrs: 32,587 individuals
   - 50–60 yrs: 21,198 individuals
   - 60–70 yrs: 4,564 individuals
   - 70–80 yrs: 500 individuals

   Combining 'less than 20' and '20-30' (to make <30yrs) results in a total of 118,146 individuals, while the older age groups (>30yrs) sum to 105,764, ensuring a relatively balanced split.
2. For the 'RACE CODE CD' feature, we retained only two categories for simplicity and relevance to the study's focus: 'BLK' (Black) and 'WHI' (White), encoded as 1 and 0, respectively. Rows of data that did not fall into these two categories were dropped.
3. The 'WEAPON I' and 'DRUG I' features were binary encoded 1 (True) and 0 (False).
4. For the target variable, SSL SCORE, scores above 250 were encoded as 1 indicating high risk, while scores of 250 and below were encoded as 0, indicating low risk aligning with [36].

In the end, we have 170,694 data instances and 12 features. The features are listed in Table 5.

*Additional feature engineering:* After getting initial results for our models, we obtained feature importance and based on the values for each feature, we decided to drop WEAPON I, DRUG I, LATITUDE I, AND LONGITUDE I from the features as they have very little contribution to the model.

**Table 5** Features selected as predictors

| Feature | Description |
|---|---|
| Age at latest arrest | The individual's age at the time of their most recent arrest |
| Victim shooting incidents | The number of times an individual has been the victim of a shooting |
| Victim battery or assault | The number of times an individual has been the victim of an aggravated battery or aggravated assault; |
| Arrests violent offenses | The number of times the individual has been arrested for a violent offense |
| Gang affiliation | Indicator if an individual has been confirmed to be a member of a criminal street gang |
| Narcotic arrests | The number of times the individual has been arrested for a narcotics offense |
| Trend in criminal activity | The trend of an individual's recent criminal activity |
| UUW Arrests | The number of times the individual has been arrested for Unlawful Use of Weapons |
| Weapon I | Is 'Yes' if at least one Weapon (UUW) Arrest in past 10 years |
| Drug I | Is 'Yes' if at least one Drug Arrest in past 10 years |
| Latitude | Latitude of Centroid of Census Tract of Arrest for the Subject's latest arrest record |
| Longitude | Longitude of Centroid of Census Tract of Arrest for the Subject's latest arrest record |

*Sensitive Features:* The sensitive features in this research are AGE and RACE. However, it should be noted that RACE is only used for bias assessment. It is not included as a predictor in this task just like the dataset provider (Chicago Police Department) excluded it.

## 3.2 Model selection and evaluation

### 3.2.1 Model selection

The models developed in our study serve a critical role in predicting the risk of being involved in a shooting incident, either as a victim or an offender. Specifically, the models predict a high risk (1) if the predicted risk score of the subject is more than 250, and a low risk (0) if the risk score is less than 250. Accurately predicting the risk of involvement in shooting incidents helps law enforcement agencies in prioritizing surveillance, intervention efforts, and resource allocation, thereby improving public safety. By mitigating age-related bias in the algorithms, we ensure that individuals are not unfairly targeted based on biased assessments.

The models selected for this task are:

- *Random forest:* Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes (classification) of the individual trees. It is known for its robustness against overfitting and ability to handle non-linear relationships.
- *Logistic regression:* Logistic Regression is a linear model used for binary classification tasks. It models the probability that a given input point belongs to a certain class. Despite its simplicity, logistic regression is highly interpretable and can serve as a strong baseline model.
- *Gradient boosting:* Gradient Boosting is an ensemble technique that builds models sequentially, with each new model attempting to correct the errors of the previous ones. This approach typically results in models that are more accurate and less prone to overfitting than individual models.

### 3.2.2 Performance metrics

To evaluate the performance of the models, we used the following metrics:

- *Accuracy:* Accuracy is the proportion of correctly predicted instances (both true positives and true negatives) out of the total instances. It provides a straightforward measure of the models' overall performance.
- *F1 Score:* The F1 score is the harmonic mean of precision and recall. It considers both false positives and false negatives, providing a balanced measure of the models' accuracy, particularly useful for imbalanced datasets.

### 3.2.3 Fairness metrics

The fairness metrics used for evaluation of the models include Equality of Opportunity Difference, Average Odds Difference and Demographic Parity. For all the fairness metrics, the prediction is considered fair is the value falls within the acceptable threshold of -0.1 to 0.1.

1. *Demographic parity:* This metric measures whether the probability of a positive outcome (e.g., being classified as high risk) is the same across different groups. In other words, demographic parity is achieved if each group has an equal chance of receiving the positive outcome, regardless of their actual proportion in the population.
2. *Equality of opportunity:* This metric specifically focuses on the true positive rate, ensuring that all groups have an equal chance of being correctly identified for the positive outcome when they qualify for it. It is a more nuanced metric that considers the accuracy of the positive predictions for each group, thereby ensuring fairness in the model's sensitivity.
3. *Average odds difference:* This metric evaluates the average difference in the false positive rates and true positive rates between groups. It combines aspects of both false positives (instances wrongly classified as positive) and true positives (correctly classified positive instances), providing a comprehensive view of the algorithm's performance across different groups. A lower value in this metric indicates a fairer algorithm, as it suggests minimal disparity in both false and true positive rates across groups.

## 3.3 *Bias* analysis and mitigation

According to [8], the Strategic Subject List dataset is highly skewed and has bias with respect to age. People over the age of 40 tend to have a low SSL score. In addition, while it is unlikely that all the people under 20 years old would be involved in a shooting, the dataset showed all of them as high risk [8]. Furthermore, *there are* 127,513 individuals on the list who have never been arrested or shot, but around 90,000 of them are deemed to be at high risk [36]. Therefore, it's critical to analyze the bias in the SSL dataset before building the algorithms.

1. The core of our analysis involves assessing the distribution of SSL scores across the demographic categories i.e., age, gender, and race. We did this analysis using histogram plots as shown in Figs. 3, 4, and 5.

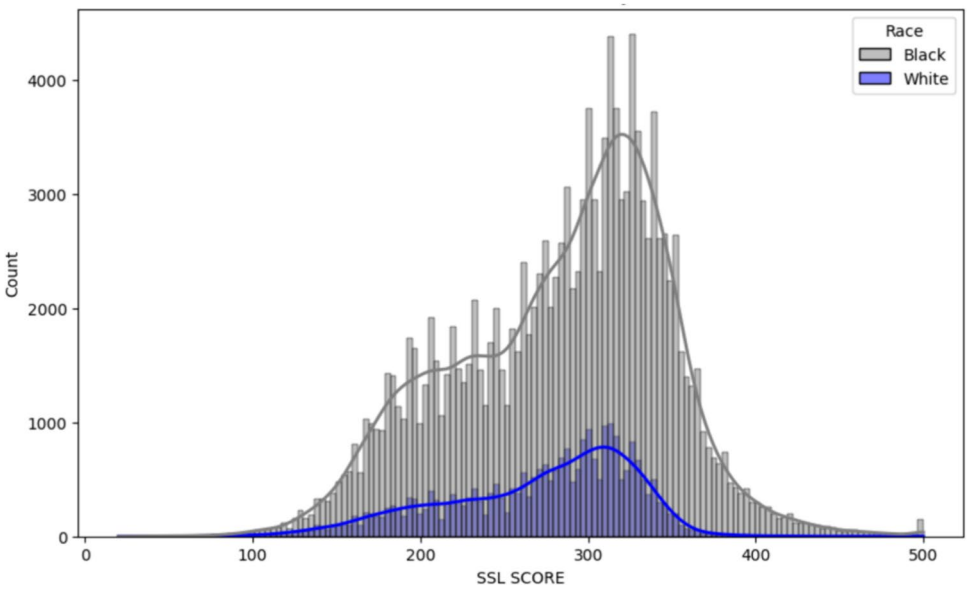


**Fig. 3** Distribution of SSL Scores by Race

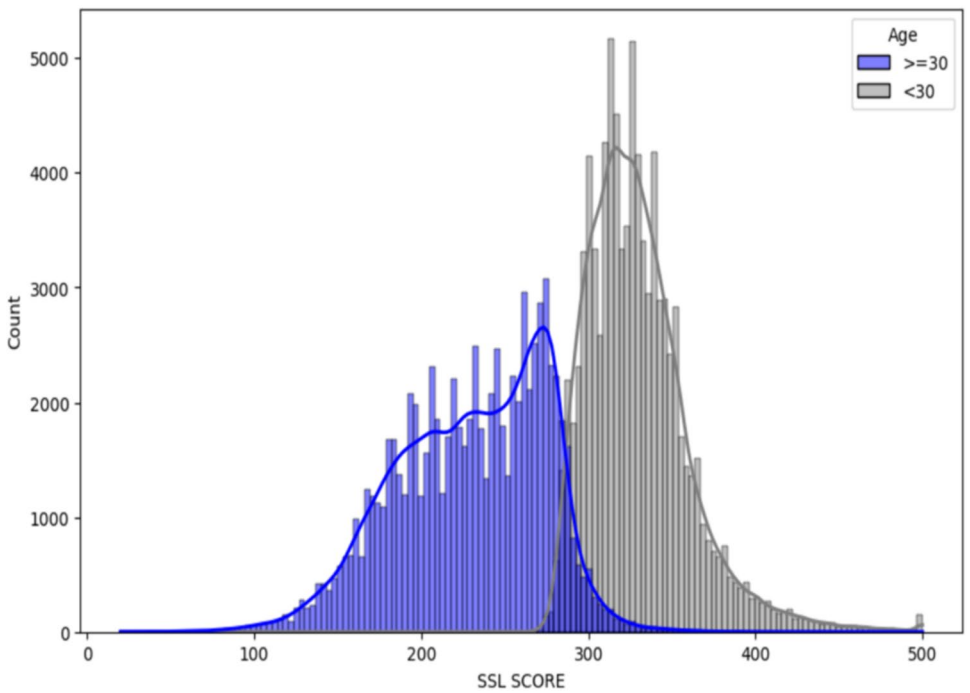


**Fig. 4** Distribution of SSL Scores by Age

2. Second, we examined a specific subset of dataset capturing individuals with SSL scores over 250 who are classified as "High Risk" despite having no prior arrest for violent offenses, no prior arrest for narcotic offenses and never been a victim of shooting incidents.

#### 3.3.1 Conditional score recalibration (CSR)

Our preliminary analysis of the dataset shows that all individuals below the age of 30 have scores higher than 250, that is, they are all marked as High Risk. This shows that the dataset is significantly skewed against young people. However, rather than implement a mitigation strategy that will

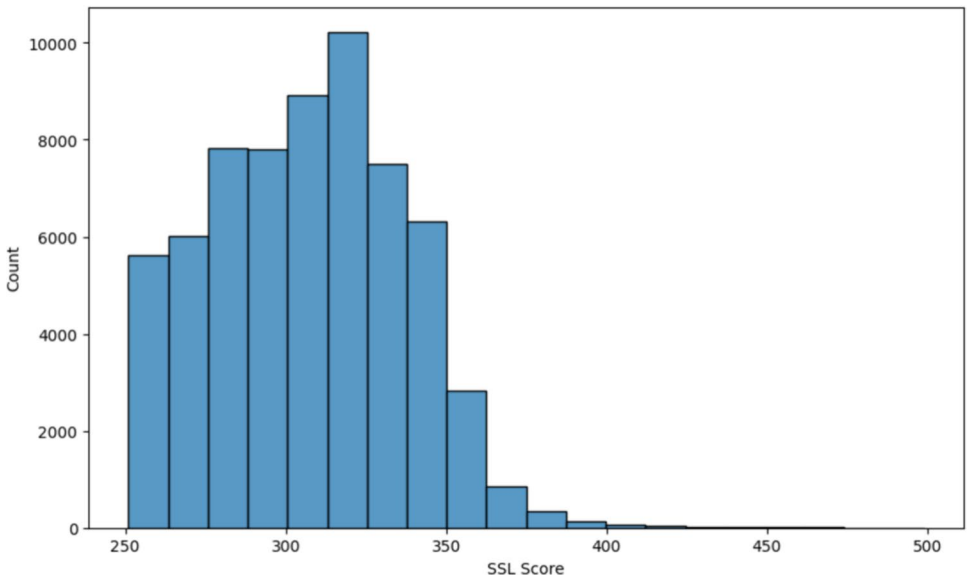


**Fig. 5** SSL Scores (> 250) for those with no violent & narcotic arrests and shooting incidents

focus on young people alone (as that would likely introduce bias against old people), we introduced a more nuanced mitigation strategy, Conditional Score Recalibration which was applied to the entire dataset, young and old. This strategy is designed to reassess and adjust the risk scores assigned to individuals within a specific subset of the dataset. Under the Conditional Score Recalibration framework, individuals originally assigned a moderately high-risk score, ranging from 250 to 350, are re-evaluated and marked as low risk if they meet ALL the following conditions:

a) No prior arrest for violent offenses
b) No prior arrest for narcotic offenses
c) Never been a victim of shooting incidents.

#### 3.3.2 Class balancing

Post-recalibration, the dataset comprised 111,117 individuals classified as low risk and 59,577 individuals classified as high risk. To address this imbalance, we employed a class balancing technique. This technique involved undersampling the majority class, which in this case was the low-risk group. By randomly selecting a subset of the low-risk individuals equivalent in size to the high-risk group, we achieved a balanced dataset.

### 3.4 Results

Figure 3 illustrates a clear racial disparity within the dataset, with a significantly higher number of black individuals compared to white. This discrepancy suggests that Black individuals are more frequently involved in arrests or encounters with the police.

Figure 4 reveals a notable difference in the age demographic within the dataset compared to the trends seen in race. First, the dataset comprises a higher proportion of younger individuals (under 30 years) compared to older individuals (30 years and above). Furthermore, a striking aspect revealed by the analysis is that all individuals in the younger age bracket are assigned SSL scores above 250, categorizing them as high risk. In contrast, the older age group shows a markedly different distribution as only a small fraction of individuals in this demographic are classified as high risk, with the majority being deemed low risk.

This disparity in risk classification between younger and older individuals raises critical considerations about the underlying factors driving these assessments. The 100% classification of younger individuals as high risk, regardless of other factors, may point to age-related biases in the Chicago Police department assessment process. This finding aligns with [36] which stated that the SSL scores assigned by the Chicago Police Department do little more than reinforce the "age out of crime" theory – a theory that suggests individuals "grow out" of crime in their 30 s. No doubt, this shows that the dataset is highly biased against young people. It also aligns with findings by [13] which noted that police actions constitute the primary contributor to model discrimination.

Figure 5 from our analysis further highlights a significant aspect of the dataset. It shows a considerable number of individuals who have not had any prior arrests for violent or narcotic offenses and have never been victims of shooting incidents, yet they have been assigned SSL scores above 250, categorizing them as high risk. This finding raises important questions about the criteria used for determining SSL scores and suggests a potential disconnect between an

**Table 6** Results obtained for Random Forest Model

| Sensitive features | Demographic parity | Equality of opportunity | Average odds difference |
|---|---|---|---|
| Race | 0.09923 | 0.08356 | 0.07679 |
| Age | 0.8517 | 0.7616 | 0.3349 |

Accuracy = 0.8314, F1 = 0.83

**Table 7** Results obtained for Random Forest after applying bias mitigation strategies

| Sensitive features | Demographic parity | Equality of opportunity | Average odds difference |
|---|---|---|---|
| Race | 0.1170 | 0.08768 | 0.03523 |
| Age | 0.3128 | 0.1521 | 0.02024 |

Accuracy = 0.9014, F1 = 0.90

individual's documented criminal history and their assessed risk level.

To determine the most suitable model for our analysis, we initially developed three different models: Logistic Regression, Random Forest, and Gradient Boosting, without incorporating any bias mitigation strategies. The performance of all three models was comparable; however, we ultimately selected the Random Forest model due to its robustness against overfitting and its ability to handle non-linear relationships in large datasets more effectively. The performance and fairness result of the Random Forest model are in Table 6.

These results reveal that the model performs fairly with respect to race, as all fairness metrics for race are within the generally accepted threshold of − 0.1 to 0.1. However, there is a significant bias regarding age indicating a pronounced disparity in the model's predictions. This high model bias against young people is not a surprise given the distribution of SSL Scores by Age in Fig. 4. After implementing the Conditional Score Recalibration and Class balancing strategies to mitigate the bias, we obtained the following results (Table 7):

The post-mitigation results show a significant improvement in the model's fairness with respect to age. The Demographic Parity metric for age decreased significantly from 0.8517 to 0.3128, and the Equality of Opportunity metric saw a reduction from 0.7616 to 0.1521. The Average Odds Difference for age also improved from 0.3349 to a much lower value of 0.02024, indicating a substantial reduction in bias. For race, the results remained largely within the acceptable threshold, with the Equality of Opportunity and Average Odds Difference metrics well within the − 0.1 to 0.1 range. The Demographic Parity metric, while slightly exceeding this range at 0.117, is still relatively close, indicating that the model maintains a fair degree of parity across racial groups.

### 3.5 Discussion

Using biased datasets without mitigation strategies can perpetuate systemic biases, leading to discriminatory practices and unjust outcomes for affected individuals. In the context of predictive policing, such biases can undermine public trust in law enforcement and worsen existing inequalities [35]. By mitigating these biases, we not only promote fairer treatment of individuals but also enhance the legitimacy and effectiveness of law enforcement practices.

In our study, we observed significant age-related biases in the Chicago Police Department's SSL dataset. From the perspective of theories of justice, particularly John Rawls' theory of justice as fairness, this unequal treatment of different age groups is problematic. Rawls' theory advocates for the arrangement of social institutions to benefit the least advantaged members of society [40]. In this context, having all 118,146 individuals under the age of 30 categorized as high risk in our dataset represents an unfair treatment. The ethical imperative, therefore, is to design and implement algorithms that do not inherently disadvantage any specific demographic group based on protected attributes like age. The CSR introduced in our study and the well-established Class Balancing technique provide a framework for achieving this goal.

The implementation of bias mitigation strategies not only improved fairness but also improved the model's overall performance. The accuracy of the model increased from 0.8314 to 0.9014, and the F1 score increased from 0.83 to 0.90. This improvement effectively dispels the commonly held notion that efforts to increase fairness invariably compromise performance. In fact, our findings are in line with recent research on policing and recidivism algorithms, such as [22] and [30], which report that there is no strict trade-off between fairness and accuracy.

Aside from fairness and accuracy, other pillars of responsible AI include privacy, interpretability or explainability, and transparency. Data collection, such as the Chicago Police Department monitoring individuals and collecting their data without consent, raises privacy concerns. However, another school of thought may argue that such surveillance is for the greater good as a secure society outweighs personal privacy concerns. Also, relying solely on model decisions is insufficient as findings by [12] and [31] show it is critical to incorporate human validation, where humans review the top features driving the model's recommendations and make the final decisions. Yet, this introduces another layer of complexity, as humans are inherently biased and may introduce additional biases. Thus, it is essential

to analyze the trade-offs between these pillars critically. Increasing transparency and explainability, for example, can help uncover the black-box nature of models, thereby enhancing trust among officers and the public [39]. However, this level of transparency might require compromising some degree of data privacy.

While our study mitigated age-related bias, it has some limitations as presented below:

1. Our SLR revealed that most existing studies on predictive policing used US datasets. In deed, the Chicago Police Department's SSL dataset reflects the specific socio-economic, cultural, and policing practices of a single city in the United States. This geographical limitation means that our findings may not be directly applicable to other regions and countries with different policing strategies, demographic compositions, and social contexts. Future research should incorporate datasets from various geographical locations to validate the findings and extend generalizability.
2. The binary classification of age groups into '$<30$' and '$>30$' simplifies the age attribute but may overlook nuances within these groups. It will be critical for future research to consider the granular details and analyze each age range to uncover additional insights.
3. Our SLR reveals that very few studies focus on biases related to socio-economic status, age, and gender compared to race. While this study has focused on mitigating age-related bias, it is critical for future studies to consider socio-economic status and gender biases as well.
4. The Conditional Score Recalibration (CSR) introduced in our study has shown promising results in reducing age-related biases. However, CSR relies on predefined criteria for recalibrating risk scores, which might not capture all dimensions of an individual's risk profile.

## 4 Conclusion

This study involves a systematic literature review on the fairness of predictive policing algorithms and mitigating age-related bias in predictive policing. From our SLR, we found out the need to expand the scope of protected attributes beyond commonly researched area like race to include gender, age, and socio-economic status. Responding to this and leveraging the Chicago Police Department's Strategic Subject List (SSL) dataset in predicting the risk of inovlvement in a shooting incident, we analyzed and mitigated age-related bias in the SSL dataset, particularly against younger individuals. Our approach involves two strategies namely Conditional Score Recalibration and Class Balancing.

We observed that the application of these strategies led to a significant reduction in age-related biases. Remarkably, this improvement in fairness did not come at the cost of accuracy. Instead, we witnessed an increase in the model's accuracy from 0.8314 to 0.9014, and an improvement in the F1 score from 0.83 to 0.90. These results challenge the prevalent notion that fairness and accuracy are mutually exclusive. Also, the study contributes to the ongoing discourse on the responsible use of AI in law enforcement, emphasizing the importance of continuously scrutinizing and refining predictive policing tools to ensure they serve the public equitably.

**Acknowledgement** The authors acknowledge the support of University College London and Prince Sultan University.

**Data availability** The dataset used for this study can be found at https://data.cityofchicago.org/Public-Safety/Strategic-Subject-List-Historical/4aki-r3np

## Declarations

**Conflict of interest** The authors declare no conflicts of interest relevant to this work.



## References

1. AbdElaziz, M., Al-qaness, M.A., Dahou, A., Ibrahim, R.A., Abd El-Latif, A.A.: Intrusion detection approach for cloud and IoT environments using deep learning and capuchin search algorithm. Adv. Eng. Softw.Softw. **176**, 103402 (2023). https://doi.org/10.1016/j.advengsoft.2022.103402
2. Abdelkader, M., Mabrok, M., Koubaa, A.: OCTUNE: optimal control tuning using real-time data with algorithm and experimental results. Sensors **22**(23), 9240 (2022). https://doi.org/10.3390/s22239240
3. Alikhademi, K., Drobina, E., Prioleau, D., Richardson, B., Purves, D., Gilbert, J.E.: A review of predictive policing from the perspective of fairness. Artif. Intell. Law. Intell. Law. (2022). https://doi.org/10.1007/s10506-021-09286-4
4. Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias, ProPublica. Retrieved April 19, 2019.
5. Berk, R.A., Kuchibhotla, A.K., TchetgenTchetgen, E.: Improving fairness in criminal justice algorithmic risk assessments using optimal transport and conformal prediction sets. Sociol. Method. Res. (2021). https://doi.org/10.1177/00491241231155
6. Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77–91). PMLR.

7. Chiao, V.: Fairness, accountability and transparency: notes on algorithmic decision-making in criminal justice. Internat. J. Law Context **15**(2), 126–139 (2019)
8. Downey, A., Islam, S. R., Sarker, M. K. (2023). Evaluating Fairness in Predictive Policing Using Domain Knowledge. In *The International FLAIRS Conference Proceedings.* https://doi.org/10.32473/flairs.36.133088
9. Fang, S., Chen, H., Khan, Z., Fan, P.: User fairness aware power allocation for NOMA-assisted video transmission with adaptive quality adjustment. IEEE Trans. Veh. Technol.Veh. Technol. **71**(1), 1054–1059 (2021). https://doi.org/10.1109/TVT.2021.3129805
10. Ferguson, A.G.: Policing predictive policing. Wash. UL Rev. **94**, 1109 (2016)
11. Heaven, W. D. (2020). Predictive policing algorithms are racist. They need to be dismantled. *MIT Technology Review*, *17*, 2020.
12. Hobson, Z., Yesberg, J.A., Bradford, B., Jackson, J.: Artificial fairness? Trust in algorithmic police decision-making. J. Exp. Criminol. (2021). https://doi.org/10.1007/s11292-021-09484-9
13. Hung, T.W., Yen, C.P.: Predictive policing and algorithmic fairness. Synthese **201**(6), 206 (2023). https://doi.org/10.1007/s11229-023-04189-0
14. Idowu, J., & Almasoud, A. (2023). Uncertainty in AI: Evaluating Deep Neural Networks on Out-of-Distribution Images. *arXiv preprint* arXiv:2309.01850.
15. Ingram, E., Gursoy, F., Kakadiaris, I.A.: Accuracy, Fairness, and Interpretability of Machine Learning Criminal Recidivism Models. In: Ingram, E. (ed.) 2022 IEEE/ACM International Conference on Big Data Computing, Applications and Technologies (BDCAT). IEEE (2022)
16. Jain, B., Huber, M., Fegaras, L., & Elmasri, R. A. (2019, June). Singular race models: addressing bias and accuracy in predicting prisoner recidivism. In *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments* (pp. 599–607). ACM.
17. Khademi, A., & Honavar, V. (2020) Algorithmic bias in recidivism prediction: A causal perspective (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, No. 10, pp. 13839–13840). AAAI Press. https://doi.org/10.1609/aaai.v34i10.7192
18. Little, R.J., Rubin, D.B.: Statistical analysis with missing data. John. Wiley. (2019). https://doi.org/10.1002/9781119482260
19. Mohler, G., Raje, R., Carter, J., Valasik, M., Brantingham, J.: A penalized likelihood method for balancing accuracy and fairness in predictive policing. In: Mohler, G. (ed.) 2018 IEEE international conference on systems, man, and cybernetics (SMC). IEEE (2018)
20. Montana, E., Nagin, D.S., Neil, R., Sampson, R.J.: Cohort bias in predictive risk assessments of future criminal justice system involvement. Proc. Natl. Acad. Sci. **120**(23), e2301990120 (2023). https://doi.org/10.1073/pnas.2301990120
21. Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., Moher, D.: The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Internat. J. Surg. (2021). https://doi.org/10.1186/s13643-021-01626-4
22. Pastaltzidis, I., Dimitriou, N., Quezada-Tavarez, K., Aidinlis, S., Marquenie, T., Gurzawska, A., & Tzovaras, D. (2022). Data augmentation for fairness-aware machine learning: Preventing algorithmic bias in law enforcement systems. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM. https://doi.org/10.1145/3531146.3534644
23. Richardson, R., Schultz, J.M., Crawford, K.: Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. NYUL. Rev. Online. **94**, 15 (2019)
24. Rodolfa, K. T., Salomon, E., Haynes, L., Mendieta, I. H., Larson, J., Ghani, R. (2020) Case study: predictive fairness to reduce misdemeanor recidivism through social service interventions. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM. https://doi.org/10.1145/3351095.3372863
25. Somalwar, A., Bansal, C., Lintu, N., Shah, R., Mui, P.: AI For Bias Detection: Investigating the Existence of Racial Bias in Police Killings. In: Somalwar, A. (ed.) 2021 IEEE MIT Undergraduate Research Technology Conference (URTC). IEEE (2021)
26. Tripathi, R.K., Jalal, A.S., Agrawal, S.C.: Suspicious human activity recognition:a review. Artif. Intell. Rev. Intell. Rev. **50**, 283–339 (2018). https://doi.org/10.1007/s10462-017-9545-7
27. Udoh, E. S. (2020, September). Is the data fair? An assessment of the data quality of algorithmic policing systems. In *Proceedings of the 13th International Conference on Theory and Practice of Electronic Governance* (pp. 1–7). ACM. https://doi.org/10.1145/3428502.3428503
28. Idowu, J. A. (2024). Debiasing Education Algorithms. International Journal of Artificial Intelligence in Education, 1–31. Springer Nature. https://doi.org/10.1007/s40593-023-00389-4
29. Ullah, W., Ullah, A., Haq, I.U., Muhammad, K., Sajjad, M., Baik, S.W.: CNN features with bi-directional LSTM for real-time anomaly detection in surveillance networks. Multimed. Tool. App. **80**, 16979–16995 (2021). https://doi.org/10.1007/s11042-020-09406-3
30. Urcuqui, C., Moreno, J., Montenegro, C., Riascos, A., Dulce, M.: Accuracy and fairness in a conditional generative adversarial model of crime prediction. In: Urcuqui, C. (ed.) 2020 7th International Conference on Behavioural and Social Computing (BESC). IEEE (2020)
31 Van Berkel, N., Goncalves, J., Hettiachchi, D., Wijenayake, S., Kelly, R.M., Kostakos, V.: Crowdsourcing perceptions of fair predictors for machine learning: A recidivism case study. Proceed. ACM Hum. Comput. Int. (2019). https://doi.org/10.1145/3359130
32 Yamni, M., Daoui, A., Karmouni, H., Sayyouri, M., Qjidaa, H., Motahhir, S., Aly, M.H.: An efficient watermarking algorithm for digital audio data in security applications. Sci. Rep. (2023). https://doi.org/10.1038/s41598-023-45619-w
33. Yen, C.P., Hung, T.W.: Achieving equity with predictive policing algorithms: a social safety net perspective. Sci. Eng. Eth. **27**, 1–16 (2021). https://doi.org/10.1007/s11948-021-00312-x
34. Gusenbauer, M., Haddaway, N.R.: Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources. Res. Synth Method. **11**(2), 181 (2020). https://doi.org/10.1002/jrsm.1378
35. Angerschmid, A., Zhou, J., Theuermann, K., Chen, F., Holzinger, A.: Fairness and explanation in AI-informed decision making. Mach. Learn Knowl. Extr. **4**(2), 556–579 (2022)
36. Posadas, B. (2017, June 26). *How strategic is Chicago's "Strategic subjects list"? upturn investigates.* Medium. Retrieved October 6, 2022, from https://medium.com/equal-future/how-strategic- is-chicagos-strategic-subjects-list-upturn-investigates-9e5b4b235a7c
37. Van Dijk, J., Nieuwbeerta, P., & Joudo Larsen, J. (2021). Global crime patterns: An analysis of survey data from 166 countries around the world, 2006–2019. *Journal of Quantitative Criminology*, 1–36.
38. Tucek, A. (2018). Constraining Big Brother: The Legal Deficiencies Surrounding Chicago's Use of the Strategic Subject List. *U. Chi. Legal F.*, 427.
39. Heinrichs, B.: Discrimination in the age of artificial intelligence. AI Soc. **37**(1), 143–154 (2022)
40. Follesdal, A.: John Rawls' theory of justice as fairness. In: Philosophy of Justice, pp. 311–328. Springer, Netherlands, Dordrecht (2014)