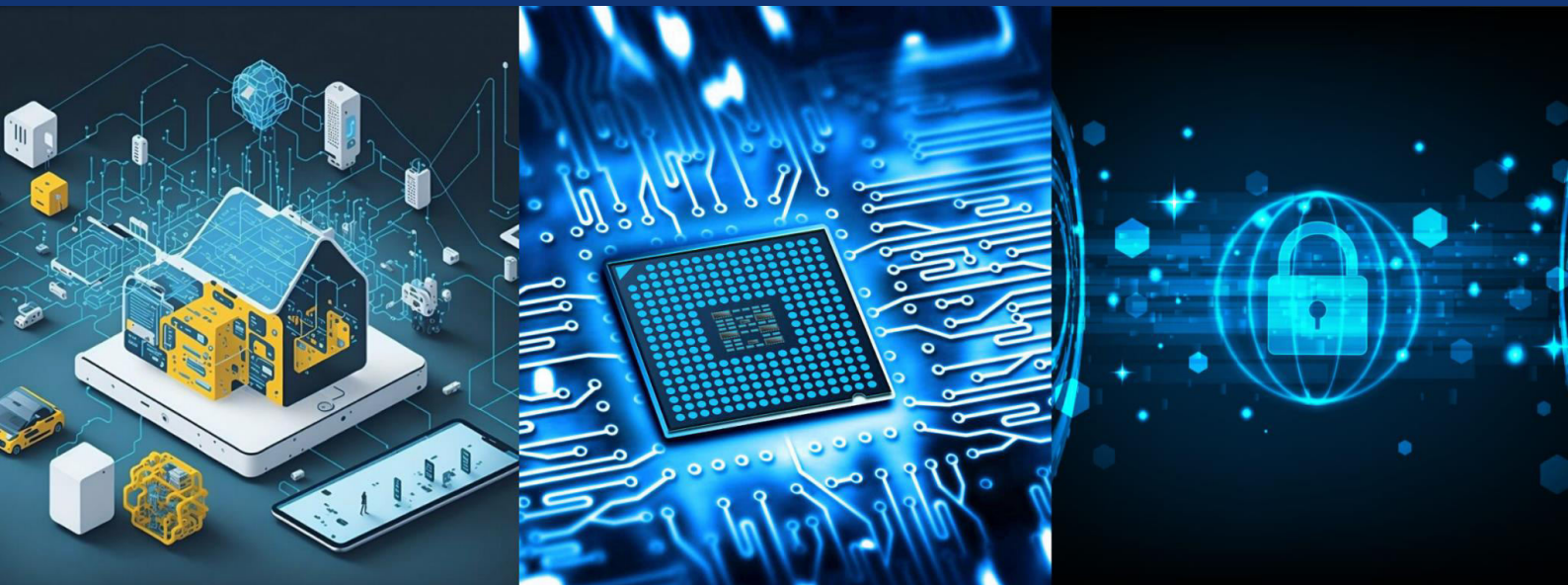
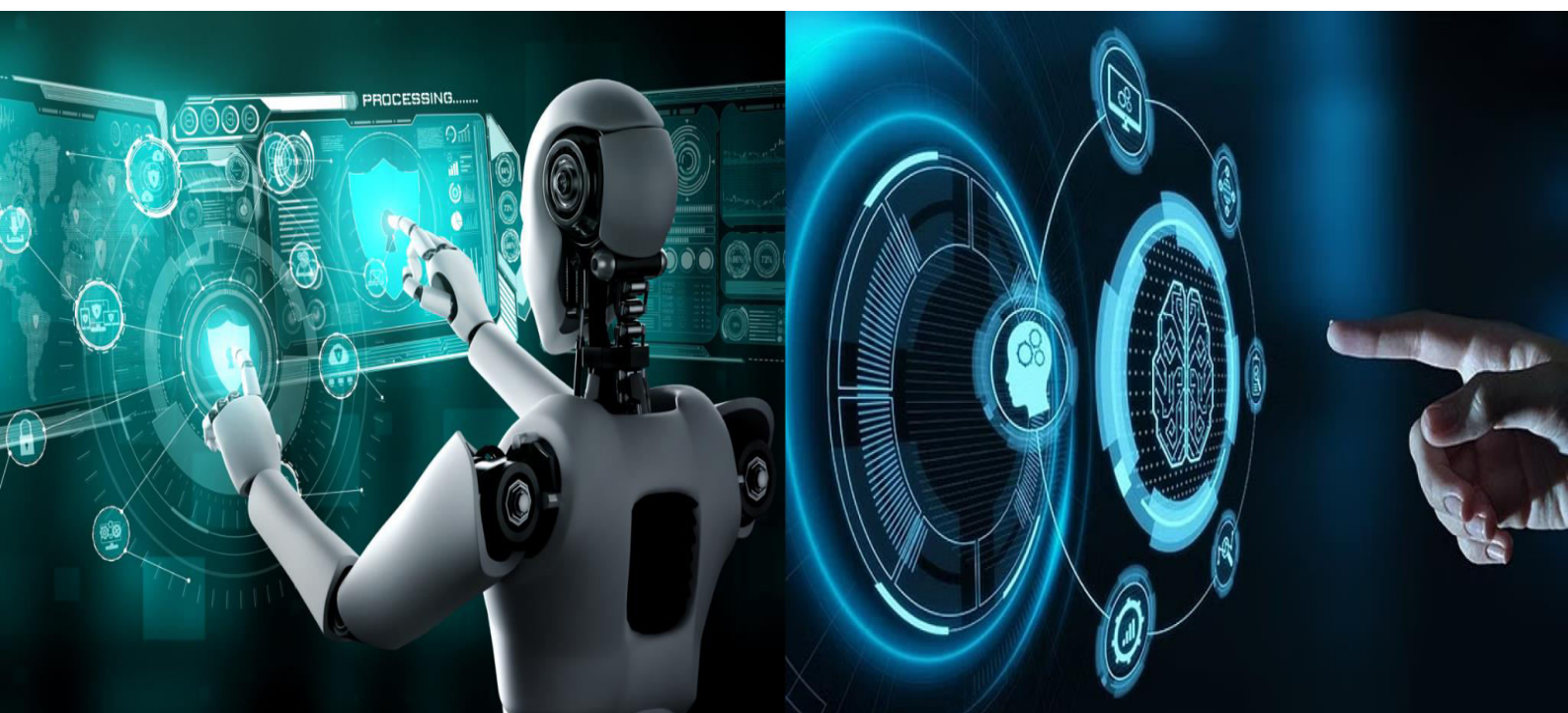




International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)





International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Dataset-Dependent Trade-offs in Bias Mitigation for General-Purpose Data Science Pipelines

Karanvir Kaur, Sahana M M

Assistant Professor, Department of Master of Computer Applications, Koshys Institute of Management Studies,
Bengaluru, India

Assistant Professor, Department of Master of Computer Applications, Koshys Institute of Management Studies,
Bengaluru, India

ABSTRACT: Machine learning models used in lending, hiring, and criminal justice can produce systematically different outcomes for different demographic groups, even when the protected attribute itself is excluded from the training features. A number of bias mitigation techniques have been proposed at different stages of the modeling pipeline — before training, during training, or after a model has already produced predictions — but there is little agreement on which stage works best, and even less evidence on whether a method that performs well on one dataset carries over to another. This paper compares three approaches on two widely used fairness benchmarks: the UCI Adult Income dataset (protected attribute: sex) and the COMPAS recidivism dataset (protected attribute: race). The methods evaluated are an unmitigated logistic regression baseline, Reweighting as a pre-processing method, and Threshold Optimizer as a post-processing method, scored on accuracy, demographic parity difference, and equalized odds difference. On Adult Income, Threshold Optimizer reduced demographic parity difference from 0.1720 to 0.0147 but raised equalized odds difference from 0.0905 to 0.3498. On COMPAS, the same method reduced both demographic parity difference (0.1613 to 0.0368) and equalized odds difference (0.1208 to 0.0204) together. Reweighting nearly halved demographic parity disparity on Adult Income at a cost of 0.4 accuracy points, but had no measurable effect on COMPAS. These results indicate that a mitigation method's fairness profile is not a fixed property of the algorithm; it interacts with dataset-specific factors such as the strength of the correlation between the protected attribute and the outcome. The paper argues for dataset-specific evaluation before deployment and for reporting more than one fairness metric, since optimizing for one can quietly worsen another.

KEYWORDS: Algorithmic fairness, bias mitigation, reweighting, threshold optimization, demographic parity, equalized odds, machine learning pipelines.

I. INTRODUCTION

Models that decide who gets a loan, whose resume gets a second look, or how a defendant's flight risk is scored are trained on historical data, and historical data carries the patterns of whoever generated it. A model does not need to be told someone's race or sex to end up treating groups differently; it can pick up the signal through features that correlate with a protected attribute, such as zip code, occupation, or prior arrest counts. The 2016 ProPublica investigation into the COMPAS recidivism tool is the case most people cite: Black defendants who did not reoffend were nearly twice as likely to be labeled high risk as white defendants who did not reoffend [7]. That single finding pushed algorithmic fairness from an academic curiosity into something regulators, journalists, and hiring managers now ask about directly. The research response has been a large number of bias mitigation techniques, grouped by where in the pipeline they intervene. Pre-processing methods adjust the training data before a model ever sees it. In-processing methods bake a fairness constraint into the training objective itself. Post-processing methods leave the trained model untouched and instead adjust its output decisions. Surveys of this literature are thorough on taxonomy but thin on practical guidance: Caton and Haas organize the field into eleven method categories and still close with five open dilemmas rather than a recommendation [1]; Pessach and Shmueli reach a similar conclusion, calling for a better understanding of which mechanism fits which scenario [2]. Hort et al. review 341 bias-mitigation publications and find the evaluation practice itself inconsistent across papers — different datasets, different metrics, different splits [3]. In short, a practitioner picking a mitigation method today has plenty of options and very little basis for choosing between them.



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

This paper runs a small, controlled comparison rather than a survey. One baseline classifier and two mitigation methods — Reweighting and Threshold Optimizer — are applied to two datasets that differ in size, base rate, and protected attribute: Adult Income (sex) and COMPAS (race). Each configuration is scored on accuracy alongside two fairness metrics, demographic parity difference and equalized odds difference, because a method can satisfy one of these while quietly violating the other. The goal is not to declare a winner. It is to show, with numbers from the same pipeline applied twice, that a mitigation method's behavior does not transfer cleanly from one dataset to another, and that this has direct consequences for how such methods should be chosen and reported in practice.

II. RELATED WORK

Fairness definitions in machine learning fall broadly into group fairness measures, such as demographic parity and equalized odds, and individual fairness measures, which ask that similar people receive similar treatment. Caton and Haas map out this space in detail and note that no single definition is uniformly preferred; the right one depends on the harm being guarded against [1]. Pessach and Shmueli organize mitigation techniques by pipeline stage — pre-, in-, and post-processing — and observe that most published comparisons pick one stage and one dataset, which makes it hard to say anything general about which stage performs best [2].

Hort et al.'s survey of 341 papers is the most direct evidence that evaluation practice in this field is not standardized: dataset choice, fairness metric, and experimental protocol vary enough between papers that results are difficult to compare [3]. Orphanou et al. take a step back and look at fairness auditing as a discipline, concluding that detection methods are more mature than certification or audit practice, and that organizations adopting these tools have little standardized guidance to follow [4]. Liu et al., writing about trustworthy AI more broadly, point out that a good deal of existing fairness work suffers from unclear or inconsistent motivation — a method is proposed and shown to reduce one metric, without much discussion of what happens to the others [5].

The two mitigation methods used in this paper both come from established, widely cited work. Reweighting was introduced by Kamiran and Calders as a way to correct for label-attribute correlation in the training set without changing any labels, by assigning each (group, label) combination a weight that would make the data look statistically independent of the protected attribute if the weights were followed exactly [9]. Threshold Optimizer builds on the equalized-odds post-processing framework proposed by Hardt, Price, and Srebro, which finds group-specific decision thresholds on top of an already-trained model's scores so that true positive and false positive rates are equalized across groups, or brought as close to equal as a single threshold shift per group allows [10]. Both are implemented in open-source toolkits — AI Fairness 360 [11] and Fairlearn [12] — that this paper uses directly rather than reimplementing from the papers.

III. METHODOLOGY

A. Datasets:

Adult Income (also called Census Income) comes from the 1994 US Census and is a standard benchmark in this literature [6]. After dropping the sampling-weight column and rows with missing values, 45,222 records remain. The protected attribute is sex, coded 0 for Female and 1 for Male. The label — whether income exceeds \$50,000 — is positive for only 11.4% of Female records against 31.2% of Male records, close to a three-fold gap that exists in the raw data before any model is trained.

COMPAS is the recidivism-risk dataset behind the 2016 ProPublica investigation, filtered according to the criteria in Larson et al.'s original methodology — restricting to defendants with a charge date within 30 days of arrest and excluding records with missing recidivism outcomes [7], [8]. After filtering, 6,172 records remain. The protected attribute is race, and the label — two-year recidivism — is positive for 45.5% of records overall, a much more balanced base rate than Adult Income, and the gap in positive rate between groups is also proportionally smaller.

B. Models and Mitigation Methods:

All three configurations use logistic regression as the underlying classifier, so that differences in outcome reflect the mitigation method rather than a change in model family.



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Baseline: logistic regression trained directly on the features, with no fairness constraint of any kind.

Reweighting (pre-processing): before training, each combination of protected-attribute value and label is assigned an instance weight so that, under those weights, the label becomes statistically independent of the protected attribute in the training set. The baseline classifier is then trained on the same features using these weights, with no other change to the pipeline.

Threshold Optimizer (post-processing): the baseline classifier's predicted scores on the test set are kept as-is, and a separate decision threshold is chosen for each group to satisfy an equalized-odds constraint. No retraining occurs; only the group-specific cutoff for a positive prediction changes.

C. Metrics:

Accuracy is reported for reference, but the paper's focus is on two difference-based fairness metrics, both of which equal zero under perfect parity and grow with disparity:

Demographic Parity Difference: the absolute difference in the rate of positive predictions between the privileged and unprivileged groups.

Equalized Odds Difference: the larger of the true-positive-rate gap and the false-positive-rate gap between groups, capturing whether the model is equally accurate for both groups rather than just equally generous.

These two metrics can move in opposite directions for the same method, which is a central finding of Section IV.

D. Experimental protocol:

Each dataset was split once into training and test partitions, and all three configurations (baseline, Reweighting, Threshold Optimizer) were evaluated on the same test split within a given dataset, so results are directly comparable within — though not necessarily across — datasets. An in-processing method, Prejudice Remover, was tested during an earlier stage of this work but is not included in the final comparison; under default hyperparameters it produced the largest accuracy loss of any method tested while barely reducing demographic parity difference. Its fairness-strength parameter was not swept in this study, so no conclusion is drawn about in-processing methods generally — this is discussed further as a limitation in Section VI.

IV. RESULTS AND DISCUSSION

Table I reports accuracy, demographic parity difference (DP Diff), and equalized odds difference (EO Diff) for all three configurations on both datasets. Figure 1 plots the same results as accuracy-versus-fairness trade-off curves alongside group-wise selection rates.

TABLE I. ACCURACY AND FAIRNESS METRICS BY DATASET AND METHOD

Dataset	Method	Accuracy	DP Diff	EO Diff
Adult Income	Baseline	0.8434	0.1720	0.0905
Adult Income	Reweighting	0.8395	0.0857	0.1368
Adult Income	Threshold Opt.	0.8259	0.0147	0.3498
COMPAS	Baseline	0.6847	0.1613	0.1208
COMPAS	Reweighting	0.6857	0.1643	0.1316
COMPAS	Threshold Opt.	0.6776	0.0368	0.0204



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

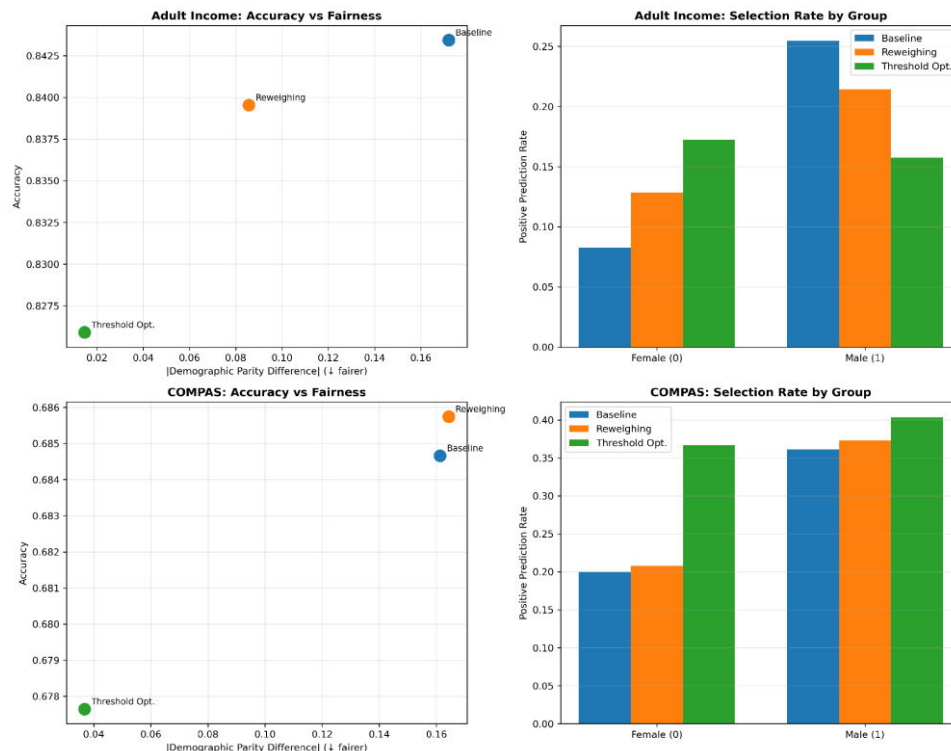


Fig.1. Accuracy vs. demographic parity difference (left) and selection rate by group (right), for Adult Income (top) and COMPAS (bottom).

A. Threshold Optimizer behaves differently on the two datasets

On Adult Income, Threshold Optimizer is the clear winner on demographic parity — it brings the difference down from 0.1720 to 0.0147, better than a ten-fold reduction — but its equalized odds difference nearly quadruples, from 0.0905 to 0.3498. Matching selection rates across a group that starts with a three-fold base-rate gap seems to require moving the decision threshold far enough that the model's error rates pull apart between groups even as its positive-prediction rates converge.

On COMPAS, the same method improves both metrics together: demographic parity difference falls from 0.1613 to 0.0368, and equalized odds difference falls from 0.1208 to 0.0204, the lowest value in the entire table. The most plausible explanation is the underlying base-rate gap. COMPAS's positive rate is closer to balanced between groups to start with, so achieving parity in selection rate does not demand as large a threshold shift, and the model's error rates do not need to diverge to get there. This is consistent with the general point that post-processing methods do not have a fixed fairness profile; their effect depends on how far the baseline model already is from parity, which is itself a property of the dataset, not the algorithm.

B. Reweighting helps on one dataset and does nothing on the other

Reweighting nearly halves demographic parity difference on Adult Income (0.1720 to 0.0857) for a 0.4-point accuracy cost (0.8434 to 0.8395), the most favorable trade-off of any method tested on that dataset. On COMPAS, it changes almost nothing: accuracy is flat, demographic parity difference moves from 0.1613 to 0.1643, and equalized odds difference moves from 0.1208 to 0.1316 — both slightly worse, not better. Reweighting corrects for a label-attribute correlation that the training data encodes; if COMPAS's bias by race runs primarily through correlated features such as prior offense counts or charge type, rather than through a direct statistical association between race and the label at the instance-weight level, reweighting the training examples would not be expected to move the needle much. This reading is consistent with the result but not proven by it — a fuller check would repeat the run across several train/test splits to rule out the result being an artifact of one particular split, since with 6,172 records COMPAS is considerably smaller than Adult Income and more sensitive to sampling variation.



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

C. What this means for choosing a mitigation method

Neither Reweighting nor Threshold Optimizer dominates the other across both datasets and both fairness metrics. A practitioner who evaluated Threshold Optimizer only on Adult Income and looked only at demographic parity would conclude it is close to a free lunch. The same practitioner looking only at equalized odds would conclude the opposite. Reporting a single fairness number, as much prior work does, hides exactly the trade-off that matters for a deployment decision. The practical implication is not that either method is bad; it is that method choice needs to be validated on the dataset and the fairness definition that matches the actual harm being guarded against, rather than imported from a benchmark result obtained on different data.

V. LIMITATIONS

Each configuration was evaluated on a single train/test split rather than averaged across multiple seeds, so the reported numbers include some amount of sampling noise that this study does not quantify. This matters most for the Reweighting result on COMPAS, where the effect size is small enough that a different split could plausibly move it in either direction; that result should be read as suggestive rather than conclusive until replicated.

Only one classifier family, logistic regression, was tested, and only two mitigation methods were carried through to the final comparison. In-processing methods were tried but excluded after Prejudice Remover underperformed under default settings; a proper comparison would sweep its fairness-strength hyperparameter rather than rely on one setting. Finally, only one protected attribute was tested per dataset (sex for Adult Income, race for COMPAS); both datasets in fact support additional protected attributes, and testing more than one per dataset would help separate dataset-level effects from attribute-level effects.

VI. CONCLUSION

Two bias mitigation methods, run through the same pipeline on two standard fairness benchmarks, produced results that do not generalize from one dataset to the other. Threshold Optimizer traded equalized odds for demographic parity on Adult Income but improved both metrics together on COMPAS. Reweighting helped substantially on Adult Income and did nothing measurable on COMPAS. The pattern that holds across both datasets is not which method wins, but that no method's fairness behavior can be assumed to carry over from a published result to a new dataset without testing it directly, and that reporting only one fairness metric can make a method look better than it is. For practitioners building on this work, the more useful takeaway may be procedural rather than a specific method recommendation: run candidate mitigation methods on your own data before deployment, and check more than one fairness definition before deciding a model is fair enough to ship.

REFERENCES

- [1] S. Caton and C. Haas, "Fairness in Machine Learning: A Survey," ACM Computing Surveys, vol. 56, no. 7, pp. 1-38, 2024.
- [2] D. Pessach and E. Shmueli, "A Review on Fairness in Machine Learning," ACM Computing Surveys, vol. 55, no. 3, pp. 1-44, 2022.
- [3] M. Hort, Z. Chen, J. M. Zhang, M. Harman, and F. Sarro, "Bias Mitigation for Machine Learning Classifiers: A Comprehensive Survey," ACM Journal on Responsible Computing, vol. 1, no. 2, pp. 1-52, 2023.
- [4] K. Orphanou et al., "Mitigating Bias in Algorithmic Systems: A Fish-Eye View," ACM Computing Surveys, vol. 55, no. 5, pp. 1-37, 2022.
- [5] H. Liu, Y. Wang, W. Fan, X. Liu, Y. Li, S. Jain, Y. Liu, A. Jain, and J. Tang, "Trustworthy AI: A Computational Perspective," ACM Transactions on Intelligent Systems and Technology, vol. 14, no. 1, pp. 1-59, 2022.
- [6] R. Kohavi, "Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid," in Proc. 2nd Int. Conf. Knowledge Discovery and Data Mining (KDD), 1996, pp. 202-207.
- [7] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine Bias," ProPublica, May 2016.
- [8] J. Larson, S. Mattu, L. Kirchner, and J. Angwin, "How We Analyzed the COMPAS Recidivism Algorithm," ProPublica, May 2016.
- [9] F. Kamiran and T. Calders, "Data Preprocessing Techniques for Classification Without Discrimination," Knowledge and Information Systems, vol. 33, no. 1, pp. 1-33, 2012.



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

- [10] M. Hardt, E. Price, and N. Srebro, "Equality of Opportunity in Supervised Learning," in Proc. 30th Conf. Neural Information Processing Systems (NeurIPS), 2016, pp. 3315-3323.
- [11] R. K. E. Bellamy et al., "AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias," IBM Journal of Research and Development, vol. 63, no. 4/5, pp. 4:1-4:15, 2019.
- [12] S. Bird et al., "Fairlearn: A Toolkit for Assessing and Improving Fairness in AI," Microsoft Research Technical Report MSR-TR-2020-32, 2020.



INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

 9940 572 462  6381 907 438  ijircce@gmail.com



www.ijircce.com

Scan to save the contact details