Communications on Applied Nonlinear Analysis
ISSN: 1074-133X
Vol 32 No. 3 (2025)

Mathematical and Statistical Tools for Ensuring Fairness in Computational Models

A. K. Chaudhary[1], K. S. Prasad[2], M. Pokhrel[3], S. Bhattarai[4], S.K. Sahani[*5], V.V. Singh[6]

Faculty of Science, Technology and Engineering, Rajarshi Janak University, Janakpurdham,Nepal

[1]Department of Management Science, Nepal Commerce Campus, Tribhuvan University, Nepal. akchaudhary1@yahoo.com

[2]Department of Mathematics, Thakur Ram Multiple Campus, Tribhuvan University, Nepal. kripasindhuchaudhary@gmail.com

[3]Department of Management, Nepal Commerce Campus, Tribhuvan University, Nepal. madanhuman@gmail.com

[4]Department of Management, Thakur Ram multiple, Tribhuvan University, Nepal. bhattaraisushil596@gmail.com

*[5]Faculty of Science, Technology, and Engineering, Rajarshi Janak University, Janakpurdham, Nepal. sureshsahani@rju.edu.np

[6]Department of Mathematics, Noida International University, India. singh_vijayvir@yahoo.com

Corresponding Authors: bhattaraisushil596@gmail.com, sureshsahani@rju.edu.np, madanhuman@gmail.com, kripasindhuchaudhary@gmail.com, singh_vijayvir@yahoo.com

Article History:

Received: *12-08-2024*

Revised: *01-10-2024*

Accepted: *11-10-2024*

Abstract:

The rising use of computational models in decision-making within healthcare, hiring, and finance underlines that fairness is of pressing concern as a means for ensuring equity, building trust, and promoting accountability. This research systematically investigates mathematical and statistical methods to identify, measure, and reduce biases of the systems. It reviews basic concepts of fairness, such as demographic parity, equality of odds, and individual fairness; methods for detecting and mitigating bias. Techniques discussed range from pre-processing adjustments to in-processing constraints to post-processing optimizations toward trading off between fairness and predictive accuracy. Application cases include the case study on ICU admissions using the MIMIC-III dataset that practically illustrates how effectiveness can be ensured with the help of fairness-aware strategies. Where there was a big gap, it now greatly closes it in demographic parity, reducing disparities in true positive rates; therefore, proving reliability for these models through sensitivity and adversarial testing. This paper bridges the divide between theory and practice by presenting a comprehensive framework towards the consideration of fairness within computational systems. The presented work identifies prospects for enhancing equity while ensuring dependability and supplements a high-growth conversation about ethical practices in AI development.

Keywords: Fairness, Bias Mitigation, Computational Models, Demographic Parity, Equalized Odds, MIMIC-III Dataset.

INTODUCTION

Throughout the past few decades, the use of computational models in decision-making systems has entered various sectors, such as hiring, financial services, criminal justice, healthcare, and more. Although these models often claim to be accurate and efficient, they are still biased in data or model that can result in unfair or discriminatory situations. Fairness of the computational models is a key part of trust, equity, and accountability in deployment. Fairness in computational domains is the process of minimizing biased results or treatment of individuals or groups, especially those who have been

historically disenfranchised. Sound statistical and mathematical techniques must be employed in order to identify, measure, and mitigate bias. Despite the considerable progress made in fair-model design, there are still several issues that need to be addressed such as contradictory definitions of fairness, conflict between the fairness and predictive performance, and constraints due to incomplete or biased data. So, to see if it's fair enough for computational models, in this study we take a look at a specific subtext of these statistical and mathematical tools. important topics including statistical methodologies for bias detection, optimization methods for fairness constraints, and methods for mitigating disparate impact. Exploring these techniques, disambiguate the layers that could lead to a useful paradigm to define fairness in computational systems.

Computational models have been used in decision-making systems in a number of industries over the last few decades, including hiring, financial services, criminal justice, healthcare, and more. These models are nonetheless biased in their data or models, which can lead to unfair or discriminating circumstances, despite their frequent claims of accuracy and efficiency. A crucial component of accountability, equity, and trust in deployment is the computational models' fairness.

In computational fields, airness refers to the process of reducing biased outcomes or treatment of people or groups, particularly those who have been historically marginalized. To detect, quantify, and reduce bias, good statistical and mathematical methods must be used. Despite the considerable progress made in fair-model design, there are still several issues that need to be addressed such as contradictory definitions of fairness, conflict between the fairness and predictive performance, and constraints due to incomplete or biased data.

No strategy can assure fairness to all parties given the inter-correlation issues [1,2], and therefore suffers the risk of not producing equitable outcomes[3,4,5,6]. Most likely other concepts of fairness [7] aspired to either the arm's or the individual’s treatment criterion, the latter being in wider use nowadays. Other considerations that aid the development of algorithms aim rather to either promote individual fairness, prevent disparate treatment or avoid disparate mistreatment of the algorithms [8, 9]. Group fairness aims at the same objectives, but this time direct the attention to decreasing the gap in the ratios of the favorable outcomes across the sub-sets of the affected population. Some of the relevant group fairness measures include difference in impact, demographic parity (statistical parity)[7,10], equality of opportunity and equal odds. There was also the norm of counterfactual fairness [11] proposed as a measure of lifting bias in causation. Unfortunately, no one has come up with a universally acceptable precision measure to use across the industry and almost all automation tasks sufficed with loose definitions of trust. Thankfully, with few exceptions, it is impossible for a single compiler to stand a chance at unifying dual principles of fairness with multilayered fairness [12, 13]across the board. Subsequently, it would be reasonable to conclude that, in its current form, setting the fairness outline will remain as shifting stance.

Fairness in computational models is a rising area of study, with manifold definitions, methodologies, and challenges. Among the most key aspects of the field are the various types of fairness, from which a basis can be derived on how to study and address the biases of models.

Individual fairness-an intuitive notion of treating similar individuals similarly-can be formalized as:

$$d(x_i, x_j) \leq \epsilon \Longrightarrow \mathrm{f}(x_i) \approx \mathrm{f}(x_j)$$

Where $d(x_i, x_j)$ measures the distance between individuals x_i and x_j $f(x)$ is the model's output, and ϵ\epsilonϵ is a threshold for acceptable similarity in treatment.

On the other hand, group fairness is about ensuring equity across different demographic groups. A and B. For example, Demographic parity: the probability of a positive outcome should be identical across groups:

$$P(\hat{Y} = 1|A) = P(\hat{Y} = 1|B)$$

If $P(\hat{Y} = 1|A) = 0.8$ and $P(\hat{Y} = 1|B) = 0.6$, demographic parity is violated. One also needs to define similarly, equality of opportunity, in which individuals with the same background (e.g., genuine label. Y=1) should have equal chances of favorable outcomes:

$$P(\hat{Y} = 1|Y = 1, A) = P(\hat{Y} = 1|Y = 1, B)$$

If $P(\hat{Y} = 1|Y = 1, A) = 0.75$ and $P(\hat{Y} = 1|Y = 1, B) = 0.85$, equality of opportunity is not achieved.

One of the key aspects of fairness research is the fairness-accuracy trade-off. This trade-off is about the notion that whenever a model is required to be fair, it has to be done at a cost to its prediction performance. For example, the measures of fairness (e.g., demographic parity or equalized odds) will most likely require altering how the model predictions or decision thresholds are set which could harm the overall accuracy of the model.

Mathematically, this cost can be expressed as a composite of the loss function.

$$L = L_{error} + \lambda L_{fairness}$$

Here, L_{error} represents the model's prediction error, $L_{fairness}$ Measures the extent of unfairness violation (e.g., differences in positive rates between groups).λ is a regularization parameter which balances accuracy and fairness. A higher λ stresses fairness, and can lead to higher fairness at a cost to the accuracy of, and a lower. λ prioritizes accuracy.

Mathematically, the binary classifier is defined by the following two confusion matrices of two demographic groups A and B:

Group	True Positive (TP)	False Positive(FP)	True Negative(TN)	False Negative(FN)	Accuracy (%)
A	80	20	90	10	85.0
B	70	30	80	20	75.0

Adding a fairness constraint, such as equalizing the false positive rates of two groups, means that there might be a need to revise the decision threshold. For instance, improving the positive prediction rate for Group B might necessitate lowering the threshold, resulting in the overall accuracy dropping slightly. In any case, the classifier is able to implement a fairer model that treats both groups better than the original model which meets the fairness objectives.

This compromise brings to light the need for a thorough situational asssessment. In some fields like hiring practice or giving credit, embracing particularity of equal opportunity fairness might translate

to preference for fairness over parity. Conversely, in situations with great risks like medical X-ray interpretation where accuracy is paramount, it might be necessary to impose large unfairness gaps which might still be a part of the process. It is a colossal challenge of striking a desire that must be balanced mathematically and understanding the impact of a chosen balancing model on society.

Mathematical and Statistical Models for Ensuring Fairness in Computational Models

1. Mathematical Model: For the formulation and resolution of issues pertaining to bias in computer systems, mathematical models offer a tangible foundation. In order for them to be modeled within optimization frameworks, fairness is defined mathematically along with particular criteria and limitations.

Therefore, definitions of fairness are the first step in mathematical modeling. Demographic parity guarantees that the likelihood of a favorable result is independent of the sensitive characteristics. The following is a simple mathematical definition of this:

$$P\left(\hat{Y} = 1 \middle| A = a\right) = P\left(\hat{Y} = 1 \middle| A = b\right)$$

Where $\hat{Y}$ represents the predicted outcome and A the sensitive attribute.

Equalized odds, another crucial idea, guarantees that mistake rates are the same for all groups. This is represented by the following formulas:

$P\left(\hat{Y} = 1 \middle| Y = 1, A = a\right) = P\left(\hat{Y} = 1 \middle| Y = 1, A = b\right)$ (True Positive Rates).

$P\left(\hat{Y} = 0 \middle| Y = 0, A = a\right) = P\left(\hat{Y} = 0 \middle| Y = 0, A = b\right)$ (True Negative Rates).

Individual fairness, on the other hand, guarantees that similar people get similar results. This is stated as

$$d\left(\hat{Y}(x_i), \hat{Y}\left(x_j\right)\right) \leq \epsilon \cdot d\left(x_i, x_j\right),$$

Where $d\left(x_i, x_j\right)$ represents the distance between two individuals in feature space.

Several popular approaches to putting fairness restrictions into practice make use of mathematical optimization techniques. Usually, the objective function is modified with a fairness penalty to make the change. The function of generalized objective is:

$$Loss = L(\omega, X, Y) + \lambda \cdot \mathrm{C}(\omega, \mathrm{A}),$$

Where $L(\omega, X, Y)$ denotes the predictive loss, $\mathrm{C}(\omega, \mathrm{A})$ quantifies fairness violations, and λ balances fairness and accuracy. Multi-objective optimization frameworks extend this to handle multiple fairness constraints simultaneously:

$$min_{\omega}\ \alpha \cdot\ L(\omega, X, Y) + \beta_1 \cdot C_1(\omega, A) + \beta_2 \cdot C_2(\omega, A)$$

Where α, β_1, β_2 are weights assigned to accuracy and fairness components

The second approach that showcases practical transparency involves adversarial debiasing, which, in practice, uses model predictions to predict sensitive information. In this case, the main model seeks to lessen bias by toying down the performance of the adversary. This leads to the following goal:

$$min_{\omega}\ L(\omega, X, Y) - \lambda \cdot \mathrm{Adv}(\omega, \mathrm{A})$$

2. Statistical Models: By providing a means of quantifying and assessing prejudice and implementing measures to mitigate it through hypothesis testing, statistical models advance equity. The divergent effect ratio, which examines the proportion of positive outcomes among various groups, is one metric for identifying bias.

$$DI = \frac{P\left(\hat{Y} = 1 \middle| A = a\right)}{P\left(\hat{Y} = 1 \middle| A = b\right)}$$

The value below 0.8 indicate potential bias.

Calibration ensures that predicted probabilities are consistent across groups, For instance:

$$P\left(Y = 1 \middle| \hat{P} = p, A = a\right) = P\left(Y = 1 \middle| \hat{P} = p, A = b\right),$$

Where Y is the actual outcome, $\hat{P}$ is the predicted probability, and A is the sensitive attribute

Hypothesis testing has become fairly prevalent for determining bias against particular groups. A t-test basically operates in much the same way, though assessing the difference of average predictions between different groups:

$$H_0: \mu_a = \mu_b, \qquad H_1: \mu_a \neq \mu_b$$

The chi-square test determines if there is a significant difference between the expected and observed outcomes for each group.

$$\chi^2 = \sum \frac{(O - E)^2}{E},$$

Where O and E are observed and expected frequencies, respectively.

The analysis of fairness is improved by regression models. To enforce fairness restrictions, regularization is added by fairness-aware logistic regression.

$$L(w) = -\sum_{i=1}^{n} [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] + \lambda \cdot \mathrm{C(w, A)}$$

3. Integration of Mathematical and Statistical Models: Because both statistical and mathematical models are used, this enables more comprehensive conclusions about fairness. While statistical techniques enable the identification and assessment of biases, mathematical optimization minimizes biases by introducing fairness constraints into model training. The model has undergone fairness tests and metrics to confirm that it has attained fairness without suffering a notable decline in accuracy. The method is designed to help computational systems make intelligent decisions without taking other factors into account.

METHODOLOGY

This is a general framework for the fairness problem in computational models, based on theories from math, stats and computer science. This framework is designed to handle biases in machine learning systems so that the models are fair and reliable. In general, the approach is broken down into a series

of steps, each based on solid theoretical underpinnings and mathematical formulations. First, fairness is defined w.r.t. various metrics such as demographic parity, equalized odds and individual fairness. These capture the many notions of fairness: from fairness across groups to fairness to similar individuals. These definitions provide the theoretical basis for bias analysis in machine learning. Then comes the actual deployment of statistical methods to find discrepancies in datasets and models. Descriptive stats, hypothesis testing and visualization are some of the techniques that give quantitative insights into the presence and extent of disparities. Independence of model outputs with respect to sensitive attributes can be obtained as an output using hypothesis tests such as chi-square tests, whereas distributional tests like Kolmogorov-Smirnov test the differences across groups. It also includes methodology on modeling and bias mitigation strategies which includes changes in pre-processing data, in-processing changes in models (perhaps adding fairness constraints) and post-processing adjustments to the outputs. In each of these stages, there are mathematical optimization techniques that include regularization that makes a trade-off between fairness and performance. Finally, robustness, generalizability and sensitivity analysis to make this framework applicable to real world.

Step 1: Problem Formulation and Defining Fairness:

The idea of fairness in computational systems draws from ethical principles of equity and justice. These concepts are then translated into specific definitions relevant to machine learning. The theoretical foundations include:

1. Demographic Parity: Based on distributive fairness, demographic parity requires that protected traits (e.g., gender or ethnicity) have no impact on outcomes. Regarding math's

$$P\left(\hat{Y} = 1 \middle| A = a\right) = P\left(\hat{Y} = 1 \middle| A = b\right)$$

Where $\hat{Y}$ represents the predicted outcome and A the protected attribute.

2. Equalized Odds: By ensuring that a model's error rates are same across groups, this statistic addresses procedural fairness. It could be expressed as

$$P\left(\hat{Y} = 1 \middle| Y = y, A = a\right) = P\left(\hat{Y} = 1 \middle| Y = y, A = b\right)$$

For $y \in \{0,1\}$, where Y represents the true label

3. Individual Fairness: Individual justice guarantees comparable results for comparable individuals, drawing inspiration from Aristotle's idea of treating like instances similarly. Applying the distance metric d

$$\boldsymbol{d}\left(\boldsymbol{f}(\boldsymbol{x_i}), \boldsymbol{f}\left(\boldsymbol{x_j}\right)\right) \leq \boldsymbol{\epsilon}, \forall_{\boldsymbol{x_i}}, \boldsymbol{x_j} \in \mathrm{X}$$

Where $f(x)$ is the prediction function

Step 2: Dataset Selection and Statistical Bias Detection:

Dataset Selection: Biased historical data, such as COMPAS for criminal justice and UCI Adult for income prediction, facilitates the exploration of these aspects of fairness. These datasets provide insights into how the systems maintain inequity and underrepresentation in certain groups.

Statistical Analysis: To find biases in datasets and models, statistical techniques are used.

1. **Descriptive Analysis and Hypothesis Testing:** It has been possible not only to study differences between distributions from different datasets but also to perform statistical applications like Kolmogorov-Smirnov and chi-square tests. For instance, it measures the chi-square test:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

2. **Causal Analysis**

The methodology is based on a causal inference methodology, an important technique for the detection of latent biases using DAGs and counterfactual simulations. This will be isolating the influence of sensitive attributes. The researchers can hence distinguish between spurious and authentic correlations. However, some challenges still remain. Whereas the process may be effective, it is somewhat intricate as it requires thoughtfulness in the underlying data structures.

3. **Visualization and Dynamic Bias Monitoring:**

Visual technologies, including fairness dashboards, incorporate real-time tracking of dynamic fairness metrics. It aids in the ongoing identification of biases in dynamic data. The problem is that there can be a lot of biases, thus it's necessary to continuously analyze these data. Although promising, the quality of the technology depends on the quality of the data. Because of this, the researchers should be extremely watchful when doing their studies and make sure that the instruments they employ are accurate and trustworthy.

Step 3: Bias Detection in Models:

Bias detection incorporates adaptive approaches and assesses fairness criteria in trained algorithms.

1. **Dynamic Fairness Metrics:** SPD and DI are dynamic throughout model evaluation, meaning they should be updated to account for changes in demographic or societal parameters over time. However, because they heavily rely on the data being examined, their efficacy may differ. Although they offer a structure for evaluating equity, the interpretation of the findings can be intricate and subtle.
2. **Parity of Error Rates**: This measure guarantees uniform rates of false positives and false negatives for all groups.

$$FPR_{A=a} = \frac{FP_{A=a}}{FP_{A=a} + TN_{A=a}}, FNR_{A=a} = \frac{FN_{A=a}}{FN_{A=a} + TP_{A=a}}$$

3. **Fairness Stress Testing** is a novel method for evaluating the robustness of models by subjecting them to simulated events, such as changes in the population or hostile inputs.

Step 4: Bias Mitigation Techniques

A more thorough explanation of these approaches, as well as the theoretical background, practical implementations, and mathematical formulations used in each of those, is provided further in this chapter. The bias mitigation techniques can balance the unfairness of computational models in three fundamental locations within the machine learning pipeline—pre-processing, in-processing, and post-processing. At each stage, statistical, algorithmic, and mathematical tools are used to ensure models are fair, equitable, and reliable.

1. **Pre-Processing** Compensated with Data Augmentation in Adversarial Training to Generate Synthetic Samples for Combating Group Imbalances

2. **In-Processing** Incorporating Regularization Based Fairness Constraints with a Dual-Objective Optimization Framework

$$min_{\theta} L_{error} + \lambda L_{fairness}(\theta), subject\ to: C(\theta) \leq \delta$$

3. **After Processing** Reinforcement learning (RL) in conjunction with threshold adjustment allows for dynamic decision optimization in real-world situations.

Step 5: Fairness-Accuracy Trade-off

Applying advanced mathematical methods to a careful examination of fairness-accuracy trade-offs is highly valued in this method.

1. **Composite Loss Functions**

The loss function integrates fairness and accuracy

$$L = (1 - \lambda)L_{accuracy} + \lambda L_{fairness}$$

A Pareto-optimal balance is achieved through adaptive tuning of λ.

2. **Analysis using game theory**:

The framework suggests a game-theoretic method by simulating different situations in which hostile players aggressively test fairness metrics. It is crucial to comprehend the weak areas in the trade-offs between accuracy and justice because these relationships can get intricate and hence confound outcomes. Despite the players' seeming independence, their tactics frequently have a significant impact on one another. Therefore, it becomes crucial to comprehend these processes in order to arrive at fair results.

Step 6: Validation and Sensitivity Analysis

Validation ensures that the strength of the fairness-preserving models is statistically and practically tested. Cross-validation and bootstrap sampling determine a particular fairness metric's stability across different subsets of data. The defined fairness metrics have their confidence intervals set as

$$CI = \hat{\mu} \pm z\frac{\sigma}{\sqrt{n}}$$

Where n is the sample size, z is the z-score for the specified confidence level, σ is the standard deviation, and $\hat{\mu}$ is the mean metric value.

Sensitivity analysis evaluates how resilient fairness is to disturbances, such as:

1. **Threshold Sensitivity**: To track changes in metrics like true positive rates between groups, decision thresholds are changed.
2. **Demographic Shifts:** To evaluate the constancy of fairness, population fluctuations are simulated.
3. **Adversarial Inputs**: To evaluate the robustness of the model, biased inputs are supplied.

A ground-breaking dynamic sensitivity score that provides flexibility in real-time scenarios combines accuracy and fairness.

$$S_{dynamic} = \alpha \cdot \frac{\partial \text{Fairness Matric}}{\partial \text{X}} + (1 - \alpha) \cdot \frac{\partial \text{Accuracy}}{\partial \text{X}}$$

Case Study: Healthcare Applications

It ensures that the fairness-aware models, with the help of the MIMIC-III dataset, provide equal ICU admission rates across racial groups while not compromising on precise mortality predictions. These tests check the dependability of the models against demographic changes, threshold changes, and adversarial perturbation in practical situations.

RESULT

These findings also support the use of fairness-aware modeling in the MIMIC-III data, which has been evaluated by numerical analysis and evidence of bias detection, mitigation, and validation in ICU admission prediction.

Bias Identification and Dataset Analysis

Disparities in ICU admissions were found by the preliminary research.

Admission Rates**:**

- Majority group $P(\hat{Y} = 1|A = majority) = 45\%$
- Minority group $P(\hat{Y} = 1|A = minority) = 30\%$
- Demographic Parity:

$$DP = \frac{P(\hat{Y} = 1|A = minority)}{P(\hat{Y} = 1|A = majority)} = \frac{0.30}{0.45} = 0.67$$

The value below 1 indicates significant bias

- Equalized Odds

True Positive Rate (TPR) gap:

$$\Delta TPR = TPR_{majority} - TPR_{minority} = 0.85 - 0.75 = 0.15$$

The TPR gap highlights unequal prediction performance across groups

Fairness-Aware Modeling

The fairness-aware methodology addressed these biases at multiple stages.

1. Pre-Processing Adjustments:

We balanced representation by assigning weights to individual samples based on group membership.

$$w_{minority} = \frac{P(A)}{P(A|Y = y)} = \frac{0.4}{0.3} = 1.33, \qquad w_{majority} = \frac{0.6}{0.7} = 0.875$$

To ensure that both groups had an equivalent impact on the parameters learned during model training, these weights were used.

2. In-Processing Constraints:

A fairness term was added to the loss function

$$L = L_{error} + \lambda \cdot L_{fairness}, \; L_{fairness} = \left|P(\hat{Y} = 1|A = minority) - P(\hat{Y} = 1|A = majority)\right|$$

Initial fairness loss: $|0.30 - 0.45| = 0.15|$

After optimization ($\lambda = 0.3$)

$$L_{fairness} = |0.42 - 0.44| = 0.02$$

3. Post-Processing Adjustments:

Each group's thresholds were changed to match the odds of admission.

$$t_{minority} = 0.5, \qquad t_{majority} = 0.6$$

Minority group $P(\hat{Y} = 1 | A = minority) = 42\%$

Majority group $P(\hat{Y} = 1 | A = majority) = 44\%$

Updated DP:

$$DP = \frac{0.42}{0.44} = 0.95$$

Performance Metrics

After applying fairness-aware techniques, the following metrics were observed

Metric	Baseline Model	Fairness Aware Model
Accuracy	85.2%	84.1%
Precision	87.3%	86.7%
Recall	82.5%	83.2%
Demographic Parity (DP)	0.67	0.95
Equalized Odds(EO Gap)	0.15	0.02

Validation and Sensitivity Analysis

After being verified mathematically, the model was tested for robustness under various circumstances.

1. Confidence Interval(CI):

Bootstrap sampling (1000 iterations) yielded,

$$CI_{DP} = 0.95 \pm 0.03, \qquad CI_{EO} = 0.02 \pm 0.01$$

These narrow intervals confirm the reliability of fairness metrics.

2. Threshold Sensitivity:

Changing thresholds by ±0.1 showed minimal fairness impact,

$$\frac{\Delta DP}{\Delta t} \approx 0.01$$

3. Demographic Shifts

Increasing minority representation (from 30% to 40%) resulted in:

$$DP = 0.93$$

4. Adversarial Testing:

When compared to the baseline, the model decreased adversarial-induced prediction bias by 65%.

Real-World Impact

We find that the fairness-aware model with the MIMIC-III dataset improves healthcare equality. While sustaining a high overall accuracy of 84.1%, the percentage of minority ICU admission increased from 30% to 42%, which means there is a shrinking gap with the majority group. True positive rate inequalities fell from 0.15 to 0.02 while demographic parity increased from 0.67 to 0.95. These results thus demonstrate how the model represents a realistic and efficient way to ensure justice in important

real-world applications such as healthcare by treating underrepresented groups more fairly without loss of predictive accuracy.

This shows that the fairness-aware approach really works in mitigating the biases8 of models, as evidenced by the MIMIC-III dataset on the ICUs. At the start, there was already unequal rates: minority admissions at 30 percent, almost half compared to the majority group at 45 percent, were admitted. It followed then that there is a gap in TPR of 0.15. After the implementation of pre-processing, in-processing, and post-processing measures, demographic parity increased from 0.67 after induction to 0.95, while the TPR gap after this period became 0.02. Validation by bootstrap sampling and confidence intervals of 0.95 ± 0.03 for demographic parity supported the hypothesis, and soundness of the findings was observed.

The more detailed analysis of the results thereafter did indeed show demographic advances and adversarial changes which were resilient. The overall models' prediction accuracy was still 84.1 percent at 42 percent of the minority share, which was an increment. These results go to support the methodology as one valid and reasonable long-term expansion which can maintain the ability to do predictive balancing of cases in real-life scenarios.

CONCLUSION

The work gives an insight into fairness within computational models by mathematically and statistically interpreting the tools for mitigating the effects of bias for fair outcomes. Fairness concepts investigated in the study include demographic parity, equalized odds, and individual fairness, and challenges considered cover issues such as trade-offs between fairness and accuracy and the lack of agreed-upon metrics for measuring fairness. Pre-processing, in-processing, and post-processing are some of the methods that, moving along a machine learning pipeline, could reduce biases without significantly affecting accuracy. We shall then see a use case for MIMIC-III dataset based fairness-aware modeling: from improving demographic parity, which moves from 0.67 to 0.95, accompanied by reductions in disparity of true positive rate between races-from 0.15 to 0.02, using race-a high value when accuracy was considered for all of the methods 84.1% were correct, proving all results produced on proposed methodology sensitive and supporting these. While the study contributes a great deal to the field, it equally acknowledges the challenges that remain resilient in achieving fairness: balancing societal and predictive objectives, adapting to evolving datasets, and overcoming biases in historical data. It emphasizes that the development of fairness-aware methodologies is a continuous process and requires interdisciplinary collaboration in building ethical AI systems. This research therefore lays a very strong foundation for the promotion of fairness in computational models and showcases the potential for achieving equitable outcomes in various real-world contexts.

REFERENCES

[1] Grgić-Hlača, N., Zafar, M. B., Gummadi, K. P., & Weller, A. (2016). The case for process fairness in learning: Feature selection for fair decision making. *NIPS Symposium on Machine Learning and the Law*, 1, 2.

[2] Garg, S., Perot, V., Limtiaco, N., Taly, A., Chi, E. H., & Beutel, A. (2019). Counterfactual fairness in text classification through robustness. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 219–226.

[3] Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review, 104*(3), 671–732.

[4] Black, E., Yeom, S., & Fredrikson, M. (2020). FlipTest: Fairness testing via optimal transport. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 111–121.
[5] Kleinberg, J., Ludwig, J., Mullainathan, S., & Rambachan, A. (2018). Algorithmic fairness. *AEA Papers and Proceedings, 108*, 22–27.
[6] Lipton, Z. C., McAuley, J., & Chouldechova, A. (2018). Does mitigating ML's impact disparity require treatment disparity? *Advances in Neural Information Processing Systems, 31*, 8125–8135.
[7] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–226.
[8] Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 259–268.
[9] Zafar, M. B., Valera, I., Rodriguez, M. G., & Gummadi, K. P. (2017). Fairness constraints: Mechanisms for fair classification. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 962–970.
[10] Calders, T., & Verwer, S. (2010). Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery, 21*(2), 277–292.
[11] Garg, S., Perot, V., Limtiaco, N., Taly, A., Chi, E. H., & Beutel, A. (2019). Counterfactual fairness in text classification through robustness. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 219–226.
[12] Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2018). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*.
[13] Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.
[14] Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review, 104*(3), 671–732.
[15] Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 797–806.
[16] Eubanks, V. (2018). Automating inequality: How high-tech tools profile, police, and punish the poor. St. Martin's Press.
[17] Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 259–268.
[18] Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.
[19] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR), 54*(6), 1–35.
[20] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). Bias and fairness in machine learning. *Nature Reviews Physics, 3*(6), 422–423.
[21] Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. *Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency (FAT)**, 149–159.
[22] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–226.
[23] Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems, 29*(NIPS 2016), 3315–3323.