

Fairness-Preserving Implementation of Machine Learning Models

Shahadat Uddin

shahadat.uddin@sydney.edu.au

The University of Sydney <https://orcid.org/0000-0003-0091-6919>

Haohui Lu

The University of Sydney <https://orcid.org/0000-0002-5338-156X>

Farshid Hajati

University of New England <https://orcid.org/0000-0002-8573-5297>

Research Article

Keywords: Machine Learning, ML Fairness, Earth Mover's Distance

Posted Date: April 18th, 2025

DOI: <https://doi.org/10.21203/rs.3.rs-6466737/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Title Page

Fairness-Preserving Implementation of Machine Learning Models

Shahadat Uddin^a, Haohui Lu^a and Farshid Hajati^b

^a School of Project Management, The University of Sydney, Forest Lodge, New South Wales, Australia

^b School of Science and Technology, University of New England, Sydney, New South Wales, Australia

Corresponding author

Dr Shahadat Uddin (PhD)

Email: shahadat.uddin@sydney.edu.au

ORCID: 0000-0003-0091-6919

Phone: +61 02 93512118

Level 2, 21 Ross Street, Forest Lodge, NSW 2037, Australia

Dr Haohui Lu (Other author)

Email: haohui.lu@sydney.edu.au

ORCID: 0000-0002-5338-156X

Dr Farshid Hajati (Other author)

Email: fhajati@une.edu.au

ORCID: 0000-0002-8573-5297

Fairness-Preserving Implementation of Machine Learning Models

Abstract

Fairness in machine learning systems is essential for building trustworthy, ethical, and socially responsible AI, particularly in high-stakes domains such as healthcare and human services. This study proposes a comprehensive fairness-preserving framework integrating data bias quantification with model-level fairness evaluation and eliminating its violation. The framework uses Earth Mover's Distance to quantify the distributional discrepancy between subgroups and the overall population, providing a statistical foundation for identifying group-level data bias. We assess fairness across five widely accepted definitions (i.e., demographic parity, equalised odds, equal opportunity, false positive rate parity, and predictive parity), each derived from the confusion matrix outcomes of ML models. The framework is empirically validated using a real-world health dataset and five commonly used supervised learning algorithms: Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, and K-Nearest Neighbours. Results show that fairness-preserving adjustments, mainly through targeted data modification, significantly reduce fairness violations with minimal impact on overall model performance. By combining data-level bias quantification with robust statistical validation, this work offers a practical and interpretable approach to implementing fairness in ML systems. The framework lays a foundation for future extensions incorporating intersectional fairness, multi-class classification, and dynamic data environments. It contributes toward the development of AI systems that are not only accurate but also equitable and accountable.

Keywords: Machine Learning; ML Fairness; Earth Mover's Distance

1. Introduction

Machine Learning (ML) models have become integral components across various sectors, including healthcare [1], finance [2], and criminal justice [3], due to their ability to derive insights, automate decision-making, and optimise [4]. ML facilitates precise diagnostics, personalised treatment plans, and predictive analytics in healthcare, significantly enhancing patient outcomes and resource management [1, 5]. ML algorithms detect fraudulent activities in finance, automate trading decisions, and improve customer service through personalised recommendations [2]. Similarly, predictive models assist in risk assessments and judicial decision-making within criminal justice, influencing sentencing and parole decisions [3]. The widespread adoption of ML systems is primarily driven by their efficiency, scalability, and predictive accuracy, allowing organisations to handle vast volumes of data effectively while reducing operational costs and improving service delivery.

However, the rapid integration of ML into critical decision-making processes has raised concerns regarding ethical implications, particularly regarding fairness, transparency, and accountability, as automated decisions increasingly affect individuals' lives [6, 7]. Decisions derived from biased ML models can perpetuate and even amplify historical biases and societal inequalities, disproportionately impacting marginalised communities. Furthermore, opaque decision-making processes inherent in many ML algorithms hinder transparency, making it difficult for affected individuals and stakeholders to understand or challenge automated decisions. These ethical concerns underscore the importance of embedding fairness considerations explicitly within ML development and deployment practices.

Fairness-aware ML models explicitly address biases that result in discriminatory outcomes for individuals or groups, especially those belonging to protected attributes, such as race, gender, or ethnicity [6]. These models incorporate specific fairness constraints into the algorithmic design and training processes, systematically identifying and mitigating biases at different stages, from data preparation to algorithmic optimisation. Fairness-aware models may utilise established fairness criteria to promote equitable outcomes and reduce bias [6]. This proactive approach helps ensure that the ML outcomes align ethically and legally with societal standards and expectations [8]. Fairness-aware ML identifies and mitigates bias throughout the implementation lifecycle, ensuring ethical responsibility, legal compliance, and public trust [9]. It helps prevent the reinforcement of harmful stereotypes and promotes equitable, trustworthy AI systems.

Despite growing recognition of fairness-aware challenges in ML, there remains a notable gap in structured methodologies for systematically evaluating and implementing fairness-preserving practices within ML workflows. Current approaches often treat data bias and model fairness as separate issues, overlooking their inherent interdependence [10]. Data bias, arising from historical prejudices, sampling errors, or measurement inaccuracies, directly influences model training and predictions, potentially resulting in discriminatory outcomes. On the other hand, model fairness involves ensuring equitable treatment and unbiased predictions for all demographic groups,

particularly those defined by protected attributes such as race, gender, or ethnicity. The lack of integrated frameworks that simultaneously address these interconnected aspects can lead to superficial or incomplete fairness solutions. This gap underscores the need for comprehensive methodologies that systematically link data bias assessment with fairness-preserving modelling practices, motivating the present research to propose an integrated framework that effectively addresses these challenges.

Previous literature has provided foundational methods for assessing and quantifying data bias using statistical distances, such as Earth Mover's Distance (EMD) [9] and Kullback-Leibler divergence [11]. EMD measures the discrepancy between two probability distributions by evaluating the minimal cost required to transform one distribution into another, effectively capturing data representation disparities among groups [9]. Similarly, Kullback-Leibler divergence quantifies the difference between distributions by measuring the relative entropy, highlighting deviations that may contribute to biased outcomes [11]. Simultaneously, fairness in ML has been examined extensively, with numerous fairness metrics and definitions proposed, including demographic parity [9], which requires equal selection rates across protected groups, and equalised odds, which ensures similar true positive rates across groups [9]. Tai et al. [12] provide a comprehensive overview of real-world tabular datasets commonly used to evaluate fairness in ML models. They delve into the relationships between protected attributes and target variables using Bayesian networks and exploratory analysis, highlighting how inherent biases in datasets can impact fairness-aware algorithm development and assessment. Saeid et al. [13] examine how default settings in ML libraries can inadvertently introduce biases and propose methods for configuring these libraries to promote fairness in model outcomes. Recent research presents a statistically validated method using k-fold cross-validation and t-tests to evaluate fairness in ML algorithms, demonstrating context-dependent fairness outcomes across datasets and models, emphasising the need for adaptable fairness definitions and enhanced fairness in ensemble approaches [14].

Despite these substantial advancements, especially in fairness-aware ML implementation, comprehensive frameworks explicitly linking the quantification of data bias directly to model fairness outcomes remain limited. There is also a notable gap in the literature regarding developing approaches that ensure the fairness-preserving implementation of ML. The absence of integrated methods that jointly consider data biases and fairness metrics hinders a holistic understanding and complete mitigation of fairness issues in real-world ML implementations, underscoring the necessity for research into unified fairness frameworks. This study addresses the gap by proposing a structured framework that integrates bias quantification methods with fairness metrics, providing a systematic approach to implementing fairness-preserving ML models. In addition to identifying data bias and model fairness, as explored in existing literature, the proposed framework enables statistically significant evaluation of these aspects to support the implementation of ML models without fairness violations.

2. The Proposed Framework

To address fairness in ML systematically, we propose an integrated framework that explicitly connects data bias assessment with model fairness evaluation and elimination. This framework recognises that fairness concerns arise at different stages of the ML pipeline and requires a structured approach to identify and address biases in data and model predictions.

2.1 Relationship Between Data Bias and Model Fairness

Data bias and model fairness are inherently connected. Biases in training data, often resulting from historical inequities, sampling imbalances, or measurement errors, directly influence model predictions and decision-making. When trained on such biased data, an ML model can perpetuate these biases, leading to unfair outcomes. For example, Amazon’s experimental hiring algorithm was found to consistently downgrade resumes that included the word “women’s,” reflecting the male-dominated industry from which its training data was sourced [15].

At the same time, even when a dataset appears unbiased, fairness issues can still arise during the training of ML models. Models trained solely to maximise accuracy may disproportionately misclassify individuals from underrepresented groups, as they tend to perform better on majority-class data. For instance, a comprehensive review highlighted that biases in deep learning models can result from various sources, including training data and algorithmic design, leading to unfair treatment of certain demographic groups [16]. This underscores the importance of incorporating fairness considerations into model development to mitigate potential biases.

Since data bias and model fairness are closely linked, both need to be assessed together rather than in isolation. Bias in the training data can carry through to the model, affecting predictions in ways that disadvantage certain groups. At the same time, even if the data is balanced, model training can introduce fairness issues due to differences in how groups are represented in learned patterns. Ensuring fairness requires identifying and addressing these issues throughout the ML pipeline rather than focusing on a single stage.

2.2 Framework Overview

The proposed framework, illustrated in Figure 1, outlines a structured approach for fairness-preserving ML implementation. The process begins with evaluating data bias before model training, ensuring that any imbalances or discrepancies in representation are identified. Methods such as EMD and Kullback-Leibler divergence can be used to quantify disparities across different demographic groups. If bias is detected, various mitigation strategies are applied, including re-weighting, resampling, or adversarial debiasing.

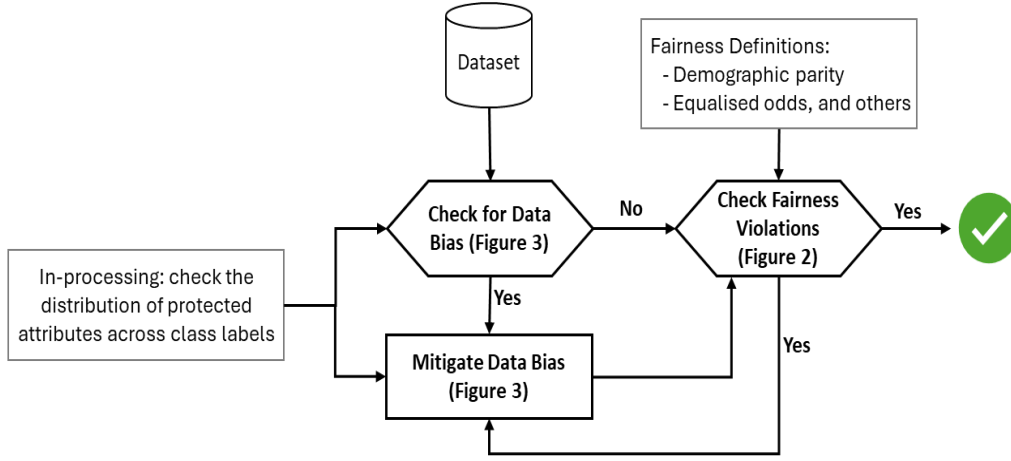


Figure 1: The proposed framework for fairness-preserving ML implementation. The framework assesses and mitigates data bias and model fairness issues to ensure equitable outcomes.

After addressing data bias, the model is assessed for fairness once training is complete. This step checks whether predictions disadvantage any protected group. Fairness is measured using different criteria, such as demographic parity, equalised odds, and equal opportunity, which compare prediction rates and error distributions across various demographic groups. If significant differences exist, adjustments such as fairness-preserving training techniques or model corrections are applied to improve fairness.

Once data bias and model fairness concerns have been addressed, the model can be deployed with greater confidence that its outcomes are equitable. The framework emphasises the need for continuous monitoring and iterative refinement, recognising that fairness is not a one-time fix but an ongoing process that requires adjustments based on real-world applications.

2.3 Data Bias and Quantification Method

Data bias arises from systematic differences in data distributions that result in unequal representation or outcomes for different demographic groups. Measuring data bias requires statistical techniques that compare distributions across these groups. EMD is a widely used method that quantifies the cost required to transform one probability distribution into another, providing insight into how data distributions vary between subpopulations. This cost is conceptualised as the amount of the distribution "mass" that needs to be moved multiplied by the distance over which it must be transported. EMD is defined as:

$$EMD = \sum_{i=1}^m \sum_{j=1}^n M_{ij} d_{ij}$$

Where M_{ij} denotes the weight transferred from the i^{th} position in the first distribution to the j^{th} position in the second distribution [9].

Kullback-Leibler (KL) divergence, another commonly used metric, measures the relative entropy between two distributions, highlighting deviations that may contribute to biased outcomes. The KL divergence from Q to P is defined as:

$$D_{KL}(P||Q) = \sum_{x \in X} P(x) \log\left(\frac{P(x)}{Q(x)}\right)$$

Detecting and quantifying bias is essential for ensuring fairness in ML models. High values of EMD or KL divergence indicate significant imbalances in representation, suggesting the need for intervention before model training. Bias mitigation at this stage typically involves adjusting the dataset through re-weighting techniques, balanced sampling strategies, or synthetic data augmentation to create a more representative and equitable distribution [6].

The proposed framework uses the bootstrap resampling technique [17] to determine the statistical significance of the transformation cost, represented by the EMD or KL value. The bootstrap method is commonly applied to estimate p -values without relying on strict parametric assumptions. For a particular subgroup, such as females, the analysis begins by isolating all instances labelled as female and calculating their corresponding EMD to evaluate statistical significance. The null hypothesis assumes no significant difference between the subgroup's EMD and that of the overall population. This is assessed by randomly reassigning female labels across the dataset using exhaustive permutations or many iterations (e.g., 10,000). The EMD is recalculated for each permutation, and the p -value is estimated as the proportion of these permuted EMD values that exceed the initially observed EMD. A p -value of 0.05 or less is interpreted as statistically significant. The p -value can be computed using the following formula:

$$p - value = \frac{\sum_i \# \{D_i \geq D_{orig}\}}{N}$$

Where, D_i is the EML or KL value from the i^{th} iteration (revised or permuted sample)

D_{orig} is the original EML or KL value

N is the total number of iterations

$\#\{\cdot\}$ denotes the count of iterations where the condition is true.

2.4 Model Fairness and Its Evaluation

Ensuring that ML models perform equitably across different demographic groups extends beyond addressing data bias. Various fairness definitions have been proposed to evaluate algorithmic equity across different population groups. Demographic parity requires the proportion of positive predictions to be consistent across subgroups [18]. Another fairness definition, equalised odds, ensures that both the true positive rate (TPR) and false positive rate (FPR) are equal across groups. In contrast, equal opportunity is a relaxation that requires only TPR parity [19]. TPR is the ratio of true positives (TP) to the sum of true positives and false negatives (TP + FN). At the same time, FPR is the proportion of false positives (FP) relative to the sum of false positives and true negatives (FP + TN). Similarly,

False positive rate parity demands equal FPRs, regardless of TPRs. Treatment equality focuses on maintaining a consistent ratio of false negatives to false positives across subgroups [3]. Lastly, predictive parity requires uniform positive predictive value (PPV) across different groups. PPV is the ratio between TP and (TP+FP). These definitions provide complementary perspectives on fairness, highlighting the trade-offs in achieving equitable outcomes across sensitive attributes.

Evaluating fairness involves testing models against these definitions using statistical methods. A recent approach includes k -fold cross-validation and statistical hypothesis testing, such as t-tests, to assess whether disparities in predictive performance are statistically significant [14]. This method provides a more robust fairness evaluation by accounting for variations across training and testing splits. As shown in Figure 2, the fairness framework divides the dataset into multiple subgroups based on the distinct values of the protected attribute. ML algorithms are trained on each subgroup using a k -fold cross-validation strategy [20]. In this method, the dataset is randomly split into k equal parts, with each fold used once as the validation set while the model is trained on the remaining $k-1$ folds. This process ensures that every instance is used for validation exactly once, yielding k independent measurements for each evaluation metric. Statistical significance is assessed by applying t-tests or ANOVA to these metric values across subgroups to determine whether fairness disparities exist. Ensuring fairness also requires consideration of protected attributes, such as race, gender, or ethnicity. A fairness-preserving approach to ML requires a comprehensive process that includes measuring bias, assessing fairness, and applying mitigation strategies to correct disparities.

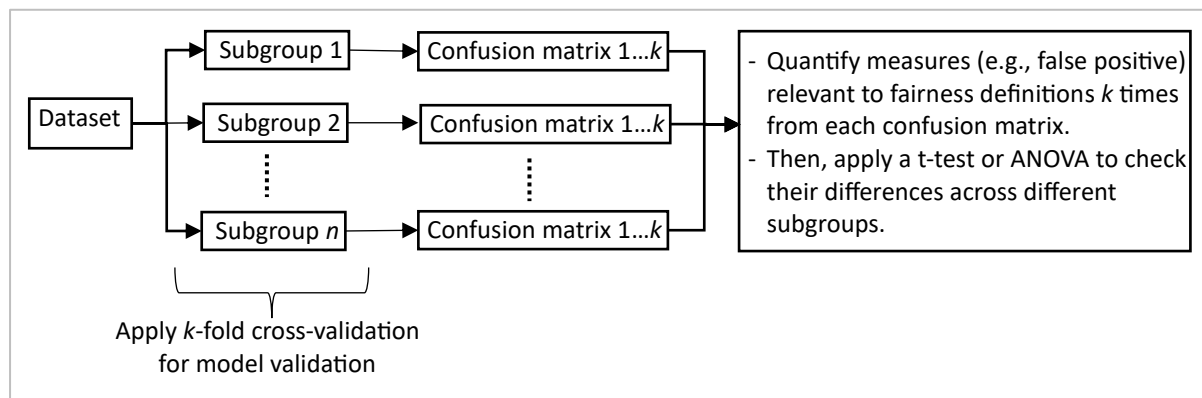


Figure 2: Fairness evaluation framework for datasets with a protected attribute

2.5 Data Bias Mitigation

Data bias mitigation is critical to developing fair and trustworthy ML models, particularly in sensitive domains such as healthcare. Once bias has been identified in the dataset using methods such as EMD or KL divergence, it becomes necessary to take corrective action before model training begins. Data bias mitigation techniques aim to reduce distributional differences between subgroups and the overall population to ensure fair representation. Approaches to mitigating data bias fall into three categories in general: pre-processing, in-processing, and post-processing methods. Pre-processing techniques,

such as reweighting or resampling instances, involve modifying the training data to reduce bias before model training [21]. In-processing approaches adjust the learning algorithm to incorporate fairness constraints or objectives during model training [22]. Post-processing methods, on the other hand, modify the model's outputs to ensure fair decision-making across different groups [19]. The choice of method depends on the context, data characteristics, and the type of fairness one aims to achieve.

We adopted a specialised in-processing approach to mitigate data bias, focusing on the distribution of the protected attribute(s) across different class labels in the dataset. A high variation of such distribution between a group of the protected attribute (e.g., females or males) and the entire sample would make the underlying EMD value higher and vice versa. As previously discussed, a higher EMD reflects a more considerable distribution divergence, signalling more substantial data bias. Since the EMD value is determined by the cost of transforming one probability distribution into another, it decreases as the corresponding class-conditional probabilities (e.g., $p(0)$ and $p(0|\text{Female})$) become more similar. Our method aims to minimise this divergence by adjusting the dataset such that these probabilities align more closely. This may involve removing certain instances from specific groups labelled with one class to reduce the distribution imbalance. The process ensures that data distributions are more equitable across demographic groups without compromising the overall data structure. Mitigation at the data level complements model fairness approaches by addressing bias at its origin. This method provides a more robust foundation for training fairness-preserving ML models by focusing on statistical distribution alignment rather than outcome balance. Figure 3 illustrates this process using an abstract dataset, showing the EMD computation and how we can mitigate data bias by lowering the EMD value.

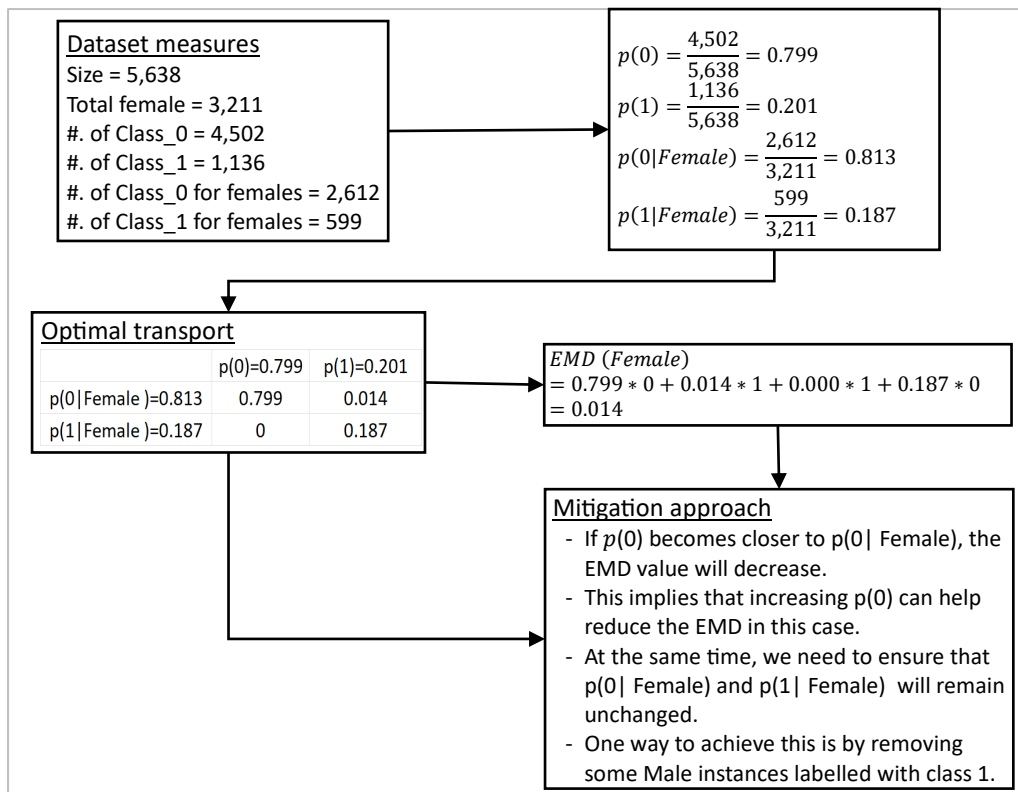


Figure 3: The Earth Mover's Distance (EMD) calculation process and data bias mitigation approach.

3. Application of the Proposed Framework

3.1 Evaluation of Data Bias

To evaluate the effectiveness of the proposed fairness-preserving framework, we applied it to a real-world dataset and assessed fairness across multiple supervised ML models. Specifically, we used the Heart Disease Prediction Dataset from Kaggle [23], which contains the demographic and medical data of 4,238 patients to predict heart disease. Of its 16 features, five represent demographic attributes, ten correspond to medical data, and the remaining one is for classification information. For simplicity of illustration, we focus solely on the EMD to demonstrate bias. We followed the steps outlined in the *Data Bias and Quantification Method* subsection to quantify EMD and its corresponding significance value.

As shown in Table 1, the dataset shows evidence of data bias related to gender, with statistically significant differences ($p \leq 0.001$). The EMD analysis revealed distributional discrepancies for male and female patients compared to the overall sample.

Table 1: Statistical assessment of gender bias for the original dataset via the Earth Mover’s Distance (EMD)

Group	EMD value	p-value	Exhibit data bias?
Female	0.0275	0.001	Yes
Male	0.0366	<0.000	Yes

3.2 Fairness Evaluation Across Multiple ML Models

Given the disparities detailed in Table 1, applying the fairness-preserving methodology allows for a comparative evaluation of fairness outcomes under different ML implementations. We considered five widely used supervised ML models: Logistic Regression (LR), k-Nearest Neighbors (KNN), Decision Tree (DT), Random Forest (RF), and Support Vector Machine (SVM). The fairness assessment was conducted based on five definitions (Demographic parity, Equalised odds, Equal opportunity, False positive rate parity and Treatment equality), ensuring a comprehensive analysis of bias propagation across models. We followed the steps outlined in the *Model Fairness and Evaluation* subsection to assess fairness violations for these five ML models. Table 2 shows the results.

Table 2: Fairness violation results for the dataset. A ‘---’ indicates no violations. The p -values for the equalised odds are for the true-positive and false-positive rates, respectively.

Fairness definition	LR	KNN	DT	RF	SVM
Demographic parity	0.005	0.026	---	0.013	0.001
Equalised odds	<0.000, <0.000	0.535, <0.000	0.056, <0.000	0.011, 0.058	0.037, 0.103
Equal opportunity	<0.000	---	---	0.011	0.037
False positive rate parity	<0.000	<0.000	<0.000	---	---
Treatment equality	0.002	---	---	---	---

Table 2 shows multiple ML algorithms exhibited violations across one or more fairness definitions. LR demonstrated the highest number of violations, failing to meet all five fairness criteria

assessed. KNN, RF and SVM each violated three fairness definitions. In contrast, DT exhibited the fewest violations, limited to equalised odds and false positive rate parity.

3.3 Fairness Improvement Through Data Bias Mitigation

Among the many data bias mitigation techniques, we followed the approach that aims to reduce the EMD for mitigating data bias in this illustration of the proposed framework. The subsection ‘*Data Bias Mitigation*’ details this approach with an illustration. Accordingly to this approach, we need to delete some male instances from class 1 to make the corresponding EMD value close to 0. After deleting 100 male instances labelled with class 1 from the dataset, the resultant EMDs showed no statistically significant disparities between groups. These EMD values become close to 0, making the dataset free from data bias, as outlined in Table 3. When we applied the fairness evaluation framework to this revised dataset, no ML algorithms revealed fairness violation issues against any fairness definitions.

Another possible alternative to make EMD close to 0 is to delete female instances from the class 0. Both $p(0)$ and $p(0| \text{Female})$ will decrease in that case. However, the latter will reduce more than the former, as it is calculated based solely on the number of female instances, whereas the former is based on the entire sample.

Table 3: Statistical assessment of gender bias after deleting 100 male instances labelled with class 1 from the original dataset via the Earth Mover’s Distance (EMD)

Group	EMD value	p-value	Exhibit data bias?
Female	0.0056	0.264	No
Male	0.0075	0.270	No

3.4 Comparison

We compared standard ML implementations with the fairness-preserving version to further assess the impact of fairness-preserving interventions. Specifically, we employed a probability distribution-based approach, which indicated that removing 100 male instances with a positive heart disease diagnosis would reduce the EMD between the subgroup and the overall outcome distributions. This data-level intervention was motivated by the framework’s objective to minimise bias before model training.

In this study, we adopted the proposed framework to predict the 10-year coronary heart disease risk. All features were normalised, and the dataset was divided using an 80:20 stratified train-test split to preserve the original class distribution. Model performance was evaluated on the test set using standard classification metrics, including accuracy, precision, recall, and F1-score, providing a consistent foundation for comparing model behaviours before and after the fairness intervention. The comparative analysis in Table 4 revealed that fairness-preserving models generally exhibited higher

overall accuracy for each ML algorithm while exhibiting adjusted precision and recall trade-offs, suggesting that the intervention redistributed error rates more equitably across demographic groups.

Table 4 Comparison of model performance before and after applying the fairness-preserving framework by removing 100 male instances with a positive heart disease outcome (TenYearCHD = 1). The framework aims to mitigate gender-based data bias. Performance is reported.

	Before applying the framework				After applying the framework			
<i>Model</i>	<i>Accuracy</i>	<i>Precision</i>	<i>Recall</i>	<i>F1-score</i>	<i>Accuracy</i>	<i>Precision</i>	<i>Recall</i>	<i>F1-score</i>
<i>LR</i>	0.8607	0.7273	0.1429	0.2388	0.8680	0.3333	0.0217	0.0408
<i>KNN</i>	0.8415	0.4091	0.0804	0.1343	0.8638	0.3684	0.0761	0.1261
<i>DT</i>	0.7527	0.1776	0.1696	0.1735	0.7809	0.1923	0.2174	0.2041
<i>RF</i>	0.8484	0.5294	0.0804	0.1395	0.8652	0.2500	0.0217	0.0400
<i>SVM</i>	0.8470	0.0000	0.0000	0.0000	0.8708	0.5000	0.0109	0.0213

The findings in Table 4 indicate that we can implement fairness-preserving ML models with minimal compromise to overall performance. This is particularly evident for the widely used accuracy metric, where all models maintained comparable or superior performance relative to their non-fairness-preserving counterparts. Standard ML models risk perpetuating and even amplifying existing societal biases without targeted mitigation. The proposed framework offers a transparent, data-centric methodology for detecting and addressing fairness concerns. It provides a practical and interpretable means of improving fairness while preserving predictive integrity by aligning subgroup distributions before model training. It is highly suitable for deployment in sensitive decision-making contexts such as healthcare. Readers can find the Python code for this application example in this GitHub link: <https://github.com/haohuilu/FairMLframework>

3. Discussion

This study presents a novel framework integrating data bias quantification with fairness-preserving ML model evaluation and eliminating its violation. The findings demonstrate that addressing data-level disparities, mainly through aligning subgroup and overall outcome distributions, can substantially improve fairness outcomes across multiple supervised ML algorithms. Notably, eliminating gender-based imbalances using a probability distribution-based approach (via EMD) led to significant improvements in fairness metrics without severely compromising predictive performance.

Importantly, the fairness-preserving modifications introduced through data-level intervention (i.e., targeted instance removal) led to performance improvements in accuracy across all models. At the same time, trade-offs in other performance metrics were observed. These trade-offs, while expected, reflect a redistribution of prediction errors in a manner that benefits overall fairness. This reinforces the idea that fairness interventions, when designed with an understanding of data structure

and statistical distribution, need not significantly undermine performance, which is a concern frequently cited in fairness-related literature.

The proposed framework is anchored in a set of well-established fairness definitions, each of which is quantified based on the four fundamental components of the confusion matrix: true positives, false positives, true negatives, and false negatives. These components are the basis for evaluating fairness-related disparities during the model testing phase. However, their values, and thus fairness outcomes, can be influenced by various underlying factors, including dataset characteristics and model architecture. For instance, as evident in the literature, dataset meta-level and statistical features of datasets significantly affect ML performance, with kurtosis showing a statistically significant negative impact on the accuracy of non-tree-based algorithms such as SVM, LR, and KNN [24]. Uddin et al. [25] reported a statistically significant performance gap between tree-based models (e.g., DT and RF) and non-tree-based models (e.g., SVM and LR), further highlighting the interplay between algorithm design and data distributional properties. These insights underscore incorporating dataset-specific characteristics when evaluating and mitigating fairness in ML models. Future research could build upon this framework by developing adaptive fairness mechanisms that dynamically account for the statistical properties of datasets and the structural nuances of different learning algorithms. Such extensions would pave the way for more robust, context-aware fairness interventions that are better suited for deployment in diverse real-world applications.

While the proposed framework offers practical value, particularly in sensitive domains such as healthcare, it has limitations. The data-level mitigation strategy may not generalise well to all fairness contexts, especially when removing instances risks losing important information. Additionally, the current framework focuses on binary classification and fairness concerning a single protected attribute (gender). Future work could extend the framework to handle multi-class settings and intersecting attributes (e.g., gender $\times$ age or gender $\times$ race), incorporating fairness notions that account for compound and systemic bias. Moreover, future studies could explore the integration of model-based fairness constraints alongside data-level interventions to offer a more comprehensive toolkit for fairness-preserving ML deployment.

5. Conclusion

This study proposed a comprehensive and statistically validated framework for fairness-preserving ML implementation, integrating bias quantification methods with multiple fairness metrics. By explicitly linking data-level disparities, measured through EMD, to fairness violations observed in model outputs, the framework offers a principled approach to identifying and eliminating discrimination in supervised learning. The empirical evaluation demonstrated that fairness-preserving interventions, particularly through targeted data-level adjustments, removed fairness violations across multiple ML models without compromising overall predictive performance.

The proposed approach advances the literature by bridging the gap between data bias quantification and practical fairness implementation. It offers a flexible and interpretable solution applicable to real-world domains such as healthcare. However, as with any data-centric intervention, the generalisability of the instance removal strategy may be limited in datasets where sample size or representativeness is a concern.

Future research could extend this framework in several important directions. These include developing methods for multi-class classification problems, incorporating intersectional fairness (e.g., considering combinations of protected attributes), and integrating model-based fairness constraints alongside data-level mitigation. Expanding the framework to handle dynamic data streams and temporal fairness would also support deployment in real-time decision-making systems. By continuing to align model performance with ethical AI principles, this work contributes to building more equitable, trustworthy, and accountable ML systems.

Declarations

Funding: The authors declare that they did not receive any funding for this study

Competing Interests: The authors declare that they do not have any competing interests

Code Availability: The Python code for the case study can be found at <https://github.com/haohuilu/FairMLframework>

Authors' contribution statement:

SU: Concept, Design, Data analysis and interpretation, Statistical analysis, Software coding, and Writing

HL: Data analysis, Software coding, Statistical analysis, and Writing

FH: Design and Writing

References

1. Uddin, S., Khan, A., Hossain, M.E., and Moni, M.A., *Comparing different supervised machine learning algorithms for disease prediction*. BMC Medical Informatics and Decision Making, 2019. **19**(1): p. 1-16.
2. Das, S., Donini, M., Gelman, J., Haas, K., Hardt, M., Katzman, J., Kenthapadi, K., Larroy, P., Yilmaz, P., and Zafar, B., *Fairness measures for machine learning in finance*. 2021.
3. Berk, R., Heidari, H., Jabbari, S., Kearns, M., and Roth, A., *Fairness in criminal justice risk assessments: The state of the art*. Sociological Methods & Research, 2021. **50**(1): p. 3-44.
4. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A., *A survey on bias and fairness in machine learning*. ACM computing surveys (CSUR), 2021. **54**(6): p. 1-35.
5. Lu, H. and Uddin, S., *Disease Prediction Using Graph Machine Learning Based on Electronic Health Data: A Review of Approaches and Trends*. in Healthcare. 2023. MDPI.
6. Caton, S. and Haas, C., *Fairness in machine learning: A survey*. ACM Computing Surveys, 2024. **56**(7): p. 1-38.
7. Gallegos, I.O., Rossi, R.A., Barrow, J., Tanjim, M.M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., and Ahmed, N.K., *Bias and fairness in large language models: A survey*. Computational Linguistics, 2024: p. 1-79.
8. Hanna, M., Pantanowitz, L., Jackson, B., Palmer, O., Visweswaran, S., Pantanowitz, J., Deebajah, M., and Rashidi, H., *Ethical and Bias considerations in artificial intelligence (AI)/machine learning*. Modern Pathology, 2024: p. 100686.
9. Pessach, D. and Shmueli, E., *A review on fairness in machine learning*. ACM Computing Surveys (CSUR), 2022. **55**(3): p. 1-44.
10. Friedler, S.A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E.P., and Roth, D., *A comparative study of fairness-enhancing interventions in machine learning*. in *Proceedings of the conference on fairness, accountability, and transparency*. 2019.
11. Ahuja, K., *Estimating kullback-leibler divergence using kernel machines*. in *2019 53rd Asilomar Conference on Signals, Systems, and Computers*. 2019. IEEE.
12. Le Quy, T., Roy, A., Iosifidis, V., Zhang, W., and Ntoutsi, E., *A survey on datasets for fairness-aware machine learning*. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2022. **12**(3): p. e1452.
13. Tizpaz-Niari, S., Kumar, A., Tan, G., and Trivedi, A., *Fairness-aware configuration of machine learning libraries*. in *Proceedings of the 44th International Conference on Software Engineering*. 2022.
14. Uddin, S., Lu, H., Rahman, A., and Gao, J., *A novel approach for assessing fairness in deployed machine learning algorithms*. Scientific Reports, 2024. **14**(1): p. 17753.
15. Rigotti, C. and Fosch-Villaronga, E., *Fairness, AI & recruitment*. Computer Law & Security Review, 2024. **53**: p. 105966.
16. Shah, M. and Sureja, N., *A comprehensive review of bias in deep learning models: Methods, impacts, and future directions*. Archives of Computational Methods in Engineering, 2025. **32**(1): p. 255-267.
17. Davison, A.C. and Hinkley, D.V., *Bootstrap methods and their application*. 1997: Cambridge University Press.
18. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R., *Fairness through awareness*. in *Proceedings of the 3rd innovations in theoretical computer science conference*. 2012.
19. Hardt, M., Price, E., and Srebro, N., *Equality of opportunity in supervised learning*, in *Advances in neural information processing systems*, D. Lee, et al., Editors. 2016, Curran Associates, Inc. p. 1-9.
20. Wong, T.-T. and Yeh, P.-Y., *Reliable Accuracy Estimates from k-Fold Cross Validation*. IEEE Transactions on Knowledge and Data Engineering, 2019. **32**(8): p. 1586-1594.
21. Kamiran, F. and Calders, T., *Data preprocessing techniques for classification without discrimination*. Knowledge and information systems, 2012. **33**(1): p. 1-33.
22. Zafar, M.B., Valera, I., Rodriguez, M.G., and Gummadi, K.P., *Fairness constraints: Mechanisms for fair classification*. in *Artificial intelligence and statistics*. 2017. PMLR.
23. Dileep and Naveen., *Logistic regression to predict heart disease*. 2025 [cited 2025 January]; Available from: <https://www.kaggle.com/datasets/dileep070/heart-disease-prediction-using-logistic-regression>
24. Uddin, S. and Lu, H., *Dataset meta-level and statistical features affect machine learning performance*. Scientific Reports, 2024. **14**(1): p. 1670.
25. Uddin, S. and Lu, H., *Confirming the statistically significant superiority of tree-based machine learning algorithms over their counterparts for tabular data*. Plos one, 2024. **19**(4): p. e0301541.