Journal of Information Systems Engineering and Management
2025, 10(35s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Dynamic Detection and Debias of Bayesian Network Classifier (3D-BN)

Fahad S. Alenazi[1*], and Khalil El Hindi[2]
[1]*Department of Computer Science, King Saud University, Riyadh, Saudi Arabia,*
Email Id: fahadsayer@gmail.com
[2]*Department of Computer Science, King Saud University, Riyadh, Saudi Arabia*
Email Id: khindi@ksu.edu.sa

ARTICLE INFO

Received: 30 Dec 2024

Revised: 19 Feb 2025

Accepted: 27 Feb 2025

ABSTRACT

Fairness in machine learning is a complex and multifaceted concept, increasingly critical in automated decision-making systems. Numerous metrics and techniques have been developed to measure and mitigate bias effectively. However, tensions often arise between different fairness notions, such as individual versus group fairness, and even among various group fairness approaches. These conflicts are typically rooted in inadequate implementation of fairness measures rather than fundamental contradictions. Additionally, failing to account for interdependencies among attributes can lead to unintended outcomes, such as those exemplified by Simpson's paradox, when focusing solely on group fairness based on sensitive attributes. This paper seeks to reconcile individual and group fairness by addressing the sources and causal dynamics of unfairness. We propose a dynamic in-process fairness enforcement method that leverages Bayesian networks and harmonizes conditional probability terms through an agnostic and symmetric objective function. Our approach aims to achieve both individual and group fairness simultaneously by applying causal path-specific bias mitigation. Moreover, it implicitly handles multiple sensitive attributes to prevent hidden redlining effects from correlated attributes and supports multi-valued attributes. A comparative evaluation of our method against related approaches using 14 real-world datasets demonstrates that our technique significantly outperforms existing fairness solutions.

Keywords: Machine Learning Fairness, Dynamic Bias Detection, Bayesian Network

INTRODUCTION

The issue of fairness in machine learning has garnered significant notice due to its impact on individuals and society at large. Numerous instances have emerged where prominent machine learning models failed to meet expectations and were deemed unjust or discriminatory [1, 2, 3]. Recently, a group of AI researchers at Apple, Amazon, Google, Facebook, IBM, Microsoft, and others founded an organization knows as Partnership on AI (PAI) which has published AI Incident Database (AIID) containing more than 1,000 AI incidents from the media and publicly available sources. Fairness issues are the most common AI incidents submitted to AIID [4]. In response, scholars have organized conferences such as the ACM Fairness, Accountability and Transparency (FAccT) conference with a focus on subjects related to Ethical and Trustworthy AI topics in automated decision making at scale [5, 6, 7].

Fairness notions proposed in the literature are usually classified in three broad areas: individual, group, and causal fairness definitions. Individual fairness aims to ensure that individuals with similar characteristics or features receive similar outcomes or treatment, while group fairness concerns with achieving equality across distinct protected groups identified by certain sensitive attributes (e.g. gender, race, or people in different age groups). On the other hand, causal fairness advocating the necessity of finding and employing causality among variables in order to really disentangle unfair impacts on decisions [8] [9] [10] [11].

However, most existing metrics and algorithms of ML fairness implicitly assume that the underlying statistical procedures or notations of fairness can be mathematically defined and deployed to create fair ML systems. This assumption generally does not involve the social and historical background of a particular field [12]. Therefore,

Journal of Information Systems Engineering and Management
2025, 10(35s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

because the applications are so different, the scope of these methods is limited. For example, while group-independent predictions make sense in employee recruitment for a certain job (when gender or ethnic factors are illegal in decision-making), this may not the case in medical applications, where gender and ethnicity can play an important role in understanding the patient's symptoms [13].

Another example is the potential conflict between individual and group fairness. For instance, while aiming to address issues related to societal inequalities on a larger scale through group fairness measures, we might produce disparate results for specific individuals belonging to those groups. The resulting biases may stem from factors such as differences in individual characteristics that are not accounted for within the broad categorization of certain populations. Nonetheless, the apparent conflict between individual and group fairness in machine learning is more of an artefact of the blunt application of fairness measures, rather than a matter of conflicting principles. In practice, this conflict between individual and group fairness may be resolved by a nuanced consideration of the sources of 'unfairness' in a particular deployment context and the carefully justified application of measures to mitigate it [14, 15].

In such cases, reconciling individual and group fairness may require the use of contextual information to identify the source of bias and which attribute should take precedence in decision-making. Causal fairness can be a useful tool for this reconciliation by understanding the root causes of unfairness and to identify ways in which individual and group fairness can be promoted simultaneously. Thus, employing a combination of statistical and causal approaches to fairness can provide a more comprehensive perspective on fair machine learning practices [16, 17, 18, 19, 20].

In this work, we proposed a framework that achieve a balance between individual and group fairness by fine-grained attribute value weighing of Bayesian network classifier to identify the causal path-specific sources of bias. The proposed method dynamically detects and debias Bayesian network classifier; we call it (3D-BN), without the explicit identification of sensitive attributes values. The aim is to improve the estimation of each conditional probability term and mitigate individual and group fairness by penalizing each attribute value with a causal influence on the model's individual and group fairness. Since we are fine tuning every attribute value, the proposed method will detect other correlated hidden sensitive attributes value if exist (red-lining effect). Moreover, our method works with binary and multi-value attributes.

We hypothesize that this approach will improve group fairness without reducing the individual fairness. We conduct extensive experiments to compare our proposed method with related (in-process) approaches on 14 ML fairness benchmark datasets. To sum up, the main contributions of our work include:

- Our inductive approach in dealing with the root cause and sources of bias in training data will improve group and individual fairness, exceedingly.
- We argue that dynamic and implicit identification of sensitive attributes will improves the model performance to extent greater than explicit identification of sensitive attributes one.
- We conduct extensive experiments to compare our proposed method with related fair ML methods on 14 benchmark datasets and compare individual and group fairness performance.

The remainder of this paper is organized as follows. In Section 2, we review related work. In Section 3, we propose our 3DBN algorithm and experimental results in detail. In Section 4, we provide our conclusions and suggestions for future research.

LITERATURE REVIEW

Bias in data-driven decision making primarily stems from the data and its collection methods. While there are numerous types of potential form of bias, statistical and representation biases are the two broad primary types of bias that frequently occur in real-world scenarios [21, 16, 18]. Statistical bias arises when the data used to train the models do not accurately represent the population while representation or historical bias occurs when the data labeling reflects existing societal biases and discrimination. This creates systematic differences among various groups and is not just limited to sample unrepresentativeness but rather affects the entire population. Therefore,

Journal of Information Systems Engineering and Management
2025, 10(35s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

some group fairness metrics are outcome comparison and not utilizing ground truth label (i.e. not impacted by data bias and focus on model bias). In addition, poor selection of features may result in a loss of important information in disproportionate ways across groups [18].

Fairness Formalization

Many works have used mathematical formulas to quantify fairness in ML. Broadly speaking, they describe some criteria that algorithms must meet to be considered "fair." Narayanan [22] and Verma and Rubin [23] provide detailed discussion of different fairness definitions. [24] also provide useful distinctions between different notions of fairness, as well as the many assumptions that justify them. In general, fairness formalization falls into three categories: group, individual and causal fairness.

Groups fairness refers to the concept that the decisions must be made independently (or conditionally independent) of group membership. For instance, the demographic parity criterion requires that predictions to be independent of the sensitive attributes [25] [26] [8]. In addition, the prediction metrics (such as accuracy, true positive rates, false positive rates, etc.) across groups must be met. For instance, the criterion of equality of opportunity requires that true positive rates are equal across groups, while the criterion of equality of odds requires that false positive rates and false negative rates are equal across groups [9] [10] [11].

Individual fairness ensures that similar individuals with respect to the prediction task are treated similarly. Feature spaces are assumed to exist in which to compute similarity, and those features will be recoverable from the data. For instance, fairness through awareness identifies a task-specific similarity metric that implies individuals who are close according to this metric are also close according to the outcome space [27] [28] [25].

Causal fairness enforces some requirements for the causal graph that generate the data and the outcomes. For instance, to ensure counterfactual fairness, it is required that there is no causal pathway from a sensitive attribute to the outcome decision [29] [26] [30]. Table 1 outlines major fairness metrics.

Table 1: Fairness Metrics

Notion	Condition
Statistical Parity (SP)	Equal acceptance rate across group
Equal Opportunity (EO)	Equal TPR across group
Equal Odds (EOdd)	Equal TPR and FPR across group
Fairness Through Unawareness (FTU)	No Explicit use of sensitive attributes
Fairness Through Awareness (FTA)	Similar individuals are given similar outcome
Counterfactual Fairness	Individual's outcome wont changed if sensitive attribute value is changed

Anti-Classification Fairness Limitations

The first definition we examine is anti-classification, which implies that classification decisions are made without considering protected attributes. Notice that anti-classification is very likely one of the first ideas that one might consider when asked for a decision system that, for example, does not discriminate against race, is not to use race attribute explicitly to make the decisions. Anti-classification approach guarantees that decisions are not based solely on group membership. Nonetheless, historical evidence demonstrates that discrimination can still occur even without reliance on protected characteristics.

An illustrative example is the use of literacy tests until the 1960s, which were apparently race-neutral but effectively discriminate against African Americans and other marginalized groups. This is because subtle correlations may

Journal of Information Systems Engineering and Management
2025, 10(35s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

exist between protected attributes and other observed or unobserved features, resulting in what Kusner et al. [30] refer to as "discrimination by proxy." For example, hobbies listed on a resume screening system could serve as an indication of gender, while zip codes associated with one's current home or birthplace might indirectly reflect race. Despite the goal of achieving fairness through unawareness, the use of proxy variables can still result in discrimination, figure 1.

In our approach, we address these issues by proposing fairness framework that can detect bias dynamically and mitigate it for any attribute including sensitive attributes while considering attributes inter-relation. As a result, our method can be generalized on different application domains that requires fairness of equal metric across group without identifying sensitive attribute explicitly.

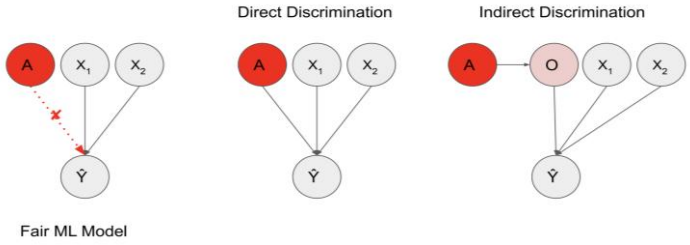


Figure 1: Discrimination by proxy for sensitive attribute (A)

Fairness Implementation

The simplest method for model fairness is to utterly strip the training information of any sensitive attributes such as demographic signals, both implicit and explicit. However, altering information might lead to loss in prediction power, and there are ways to incorporate into the model's training design some fairness measures without sacrificing model performance. These different approaches to increase fairness and mitigate biases in the Machine Learning literature in general are organized into widely accepted frameworks of pre-processing, in-processing, and post-processing methods [31, 32, 33, 8, 34]:

- **Pre-processing methods**

  Pre-processing methods focus on adjusting the training data distribution to balance the sensitive groups. These methods transform the data before the machine learning models learn from it. Examples include reweighting and resampling, as well as more complex methods like optimized data transformation which reduces bias and the predictability of the protected or sensitive attribute [35] [36] [37]

- **In-processing methods**

  In-processing methods take fairness directly into consideration during model design to induce intrinsically fair models and fundamentally mitigate fairness issues in outputs and representations. They constrain machine learning models while they learn. These methods can be categorized into explicit and implicit mitigation methods based on where the fairness is achieved in the model. Explicit methods directly incorporate fairness metrics in training objectives, while implicit methods focus on refining latent representation learning [38, 39, 40, 41]

- **Post-processing methods**

  Post-processing methods calibrate the prediction results after model training. They make predictions from a black-box machine learning model fair in the post-processing stage. These methods are versatile for correcting bias in machine learning systems that are already used in production, avoiding expensive retraining [11] [42]

Following [39, 40], in this study, we design (in-process) regularization strategy to the training objective that quantifies the degree of bias for the system training to maximize prediction performance while minimizing discrimination. We show that these discriminations are often due to scarce or skewed historical bias data for

Journal of Information Systems Engineering and Management
2025, 10(35s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

underrepresented minority or protected group, where one group has more training examples for some outcome than another.

Related Fair Bayesian Classifiers

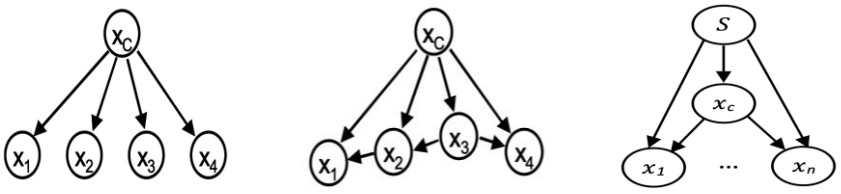
Recent research efforts aim to improve Naïve Bayes fairness and focus on static fairness, which requires identifying the set of sensitive attributes in advance and tailor the objective function of the model to improve its fairness. As a result, the model is limited to application domains in which the sensitive attribute is known. Two-naive-Bayes (2NB) [43], and N-naive-Bayes (NNB) [44] algorithms partition the data based on sensitive attribute value and use in-processing routine to enforce statistical independence of the label and the sensitive attribute. For example, given a training set of labeled instances, the algorithm partitions the data based on the sensitive attribute value and trains a separate naive Bayes sub-estimator on each of the subsets then uses the related model based on the query instance sensitive value.

In [45], a proposed fair Bayes-optimal classifier incorporates a post-processing technique to mitigate NB classifier unfairness over protected groups and achieves a better fairness-accuracy tradeoff. First, they define a threshold for NB classifier as the intersection points of any vertical line with two conditional probability densities of Y = 1 for the two groups values of sensitive attribute. Then, the classifier balances the thresholds for the two groups, by increasing the threshold for the group with a higher proportion classified as "1" and thus bringing the proportions classified as "1" closer. The procedure is similarly applied for "0" class.

The Bayesian network (BN) can be classified as a transparent model unlike other black box models, such as neural networks, and can help identify the sources of bias by explicitly modeling the relationships and the causal influence between input variables and output variables. Despite that BN is still susceptible to data biases, the causal relations between variables can help identifying any factors that are unfairly influencing the model's predictions, and to adjust the model accordingly.

The determination of interdependencies among random variables in a domain becomes exponentially complex when the Bayesian Network's structure is unknown. The process of inference within Bayesian Networks (BN) is not only NP-hard but also NP-hard for approximate inference to achieve a fixed level of accuracy [46]. The absence of knowledge regarding the correlation between random variables impedes the reduction of the joint probability distribution. In the context of a BN classification model, it becomes imperative to ascertain the joint probability distribution conditioned on the class. To address this challenge, various simplified Bayesian structures tailored for classification tasks have been proposed. Notable examples include naïve Bayes and Tree-augmented naïve Bayes, as depicted in Figure 2. These structures play a significant role in defining the classification model, which can be expressed by the following formula:

$$P(c|a_1 \dots a_n) = P(c) \prod_{i=1}^{n} P(a_i|parent(a_i) \wedge c)) \quad (1)$$



(a) Naïve Bayes (b) TAN (c) 2NB

Figure 2: Naïve Bayes, Tree- Augmented naïve Bayes and Two-naive-Bayes (2NB)

DYNAMIC DETECTING AND DEBIAS BAYESIAN NETWORK (3DBN)

In this work, we aim to improve individual and group fairness simultaneously while employing causal fairness. Precisely, we will first balance the model performance metric (i.e. fpr, tpr … etc.) across each attribute value and for each outcome to mitigate the root cause of bias and enforce overall equal metric across group. Making fairer outcome for each attribute value will assure group fairness represented by the subset of sensitive attribute value but

Journal of Information Systems Engineering and Management
2025, 10(35s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

might affect model's performance and individual fairness which we will tackle in the second step. Moreover, by enforcing fair outcome for each attribute value, we will assure mitigating redlining of correlated hidden sensitive attributes. Secondly, we will promote discrimination by boosting high predictive attribute value (and more importantly hidden high predictive attribute value) to improve its predictive power influence on classifying the correct target class. Thus, we will improve the classification performance (the mapping function between input features and outcome) which in turn improves individual fairness by classifying similar individual to similar outcome.

Improve Group Fairness

To improve fairness across groups, we will enforce fairness across each attribute value by measuring its instances' performance metric across outcomes. Then, we will decide if we should increase or decrease its conditional probability terms. For example, if the training examples that contain the attribute value has lower true positive rate compared to false positive rate, then we should increase its conditional positive class probability terms. However, we should also compare the true negative and the false negative rates in the same manner.

Now the performance metrics could conflict each other and to priorities which metric is more important than others on deciding the update direction, we have two options: first, we can make it as user input to decide which metric is more important (i.e. SP is more important than EO, and therefore, true positive is more important than true negative). The second option is to incorporate all four values of the confusion metric (tp, fp, tn, and fn) equally using a composite score such as Matthews Correlation Coefficient (MCC). MCC is Symmetric where no class is more important than others even if one class is disproportionately under- (or over-) represented.

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (2)$$

Improve Individual Fairness

The update amount of attribute value probability terms will play a key role for maintaining individual fairness while improving group fairness. Thus, we want to increase discriminative attribute value predictive power while balancing its influence across different outcomes. In our proposed method, we delineate three scenarios concerning the distribution of conditional probability terms for attribute values. In the first scenario, a certain potentially discriminatory attribute value, denoted as Da_{ij}, might be under-represented in the training data. Consequently, the conditional probability term $P(Da_{ij}|C)$ for both the ground truth label and other class labels will be substantially small due to inadequately represented data and weak correlation between the ground truth label and other classes, respectively. We refer to such attribute value as hidden discriminatory attribute values, leading to incomplete information and, consequently, causing underfitting which will generate high misclassification rate in both training and testing data. Therefore, it is imperative to significantly enhance attribute values with small conditional probability terms $P(Da_{ij}|C)$ for both the ground truth label and other class labels.

In the second scenario, some potential discriminatory attribute value might be under-sampled due to class-imbalanced dataset, where numerous examples belong to one or more major classes, and only a few belong to minor classes. In this scenario, certain discriminatory attribute value (Da_{ij}**)** may be considered as noise examples leading to an over fitting problem due to bias toward major classes compared to the rare classes. It is crucial to distinguish these examples from those in the third scenario that are strongly correlated with both classes. The former examples are affected by under-sampling problem, which is very common in real world application, whereas the later should be regarded as redundant information with no predictive power given their relatively highly correlations with different classes and being unaffected by scarce data issues.

To address these three different scenarios, we can employ a disproportional probability term boost for attribute values using the harmonic average. This choice is motivated by the harmonic average's sensitivity to smaller values, making it apt for the task [47]. Precisely, for scenario 1, the complement harmonic average (1- harmonic average) would be large and the update size would be large if both the $p(a_i|c_+ \wedge parent(a_i))$ and $p(a_i|c_- \wedge parent(a_i))$ were to be small. In this context, c_+ and c_-represent the different outcome classes. Similarly, for scenario 2 of skewed

Journal of Information Systems Engineering and Management
2025, 10(35s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

data, the complement harmonic average would be relatively large, and the update size would be large if either $p(a_i|c_+ \wedge parent(a_i))$ or $p(a_i|c_- \wedge parent(a_i))$ were to be small. Finally, in scenario 3, the complement harmonic average would be small, and the update size would be small if both $p(a_i|c_+ \wedge parent(a_i))$ and $p(a_i|c_- \wedge parent(a_i))$ were to be large. Thus, we calculate the update weights for $p(a_i|c_+ \wedge parent(a_i))$ and $p(a_i|c_- \wedge parent(a_i))$ of each attribute value using Eqs. (3).

$$W_i = \frac{\eta}{t}\left(1 - 2/\left(\frac{1}{p_t(a_i|\, c_{Positive} \wedge\ parent(a_i))} + \frac{1}{p_t\left(a_i|\, c_{Negative} \wedge\ parent(a_i)\right)}\right)\right) \quad (3)$$

Here, (η) is a learning rate between zero and one, and (t) is the iteration (epochs) number and used as weight decay.

We argue, that applying this heuristic rule aligns with the evidence observed in the training data. In scenarios 1 and 2, the model's misclassification of training examples is attributed to underfitting and overfitting, respectively. It is reasonable to assume that there is insufficient data to support the accurate classification of these training instances. Despite the global non-linearity in the relationship between attribute values and class prediction, our proposed method establishes a local linear relationship for discriminative attribute values. This localized approach is robust enough for a Bayesian classifier to discern and significantly enhance potentially hidden discriminative attribute values, thereby augmenting its predictive capability.

Improve Causal Fairness

To address the causal fairness, we may use causal influence quantification to answer fairness questions. For instance, consider a hypothetical scenario where the loan rejection rates are higher for women compared to men. This discrepancy could be attributed to the fact that women might be applying for loans to establish businesses with higher inherent risks. In contrast, men may predominantly seek loans for businesses with lower risk profiles, consequently leading to lower rejection rates. This outcome could be explained by hidden variables such as differences in business risk, among male and female applicants which influenced acceptance rates and ultimately

resulted in reversing the overall trend observed in the data. From a causal perspective, what is important is the direct impact of the protected attribute (in this case, gender) on the decision (loan approval), which cannot be attributed to any other factor such as business risk.

In our proposed method (algorithm 1) where we employ Bayesian network (TAN) which require that a causal link be established between attributes and the decision, we are able to carefully assess unfairness at a deeper level by balancing the causal path for each attribute value to the decision. For instance, in our previous example, we will find that business risk has more influence on the decision. Thus, the gender value given the business risk as parent might has no bias. Therefore, improving causal path represented by the conditional probabilities will impact positively individual and group fairness metrics than arbitrary balancing attribute value independently from other correlated attributes.

Algorithm 1: Pseudocode for a probability-balancing routine to enforce statistical parity

Build initial BN classifier
while training MCC improve and *t* < T **do**
for each attribute value, *ai*, given the class value C and other parent of *ai* (if exists)
Calculate the **tpr** and **tnr** of training subset containing *ai*
if tpr <= tnr **then**
$\boldsymbol{p_{t+1}(a_i|c_+ \wedge parent(a_i)) \mathrel{+}= W_i}$
$\boldsymbol{p_{t+1}(a_i|c_- \wedge parent(a_i)) \mathrel{-}= W_i}$
else
$\boldsymbol{p_{t+1}(a_i|c_- \wedge parent(a_i)) \mathrel{+}= W_i}$

Journal of Information Systems Engineering and Management
2025, 10(35s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

$\boldsymbol{p_{t+1}(a_i|c_+ \wedge parent(a_i))} -= \boldsymbol{W_i}$

end if

Let $\boldsymbol{p_{t+1}(a_i|C \wedge parent(a_i))}$ be equal to $min(max(\boldsymbol{p_{t+1}(a_i|C \wedge parent(a_i))}, 1e\text{-}5), 1\text{-}1e\text{-}5)$

Let $t = t + 1$

end while

EXPERIMENT SETUP AND RESULT

In various application scenarios, the assessment of fairness in machine learning models can vary based on the desired classification performance criteria. For instance, in pretrial risk assessments, achieving equal false positive rates across all groups may be prioritized, as it is often more acceptable to let a guilty person go free than to wrongfully incarcerate an innocent individual. Conversely, in loan approval systems, one may prefer a decision-making process where false negative rates are equal, ensuring that individuals deserving of loans (positive class) are not disproportionately denied (negative class) based on a specific sensitive attribute value. Furthermore, depending on the specific application and the associated costs of misclassification, disparate mistreatment may be assessed using false discovery and false omission rates instead of traditional false positive and false negative rates.

In our experiment, we will make our metric more generic, and we will use three Equal Metrics Across Groups as fairness criterion for group fairness Namely, SP, EO, and EOdd. In addition, we will use consistency to measure individual fairness. We also, average accuracy and F-Score for each sensitive attribute value to have fairer metrics for the models and not marginalizing them over sensitive attribute. The counterfactual fairness has limitation where it doesn't address other hidden sensitive attributes and/or other correlated attributes, thus, we opt not to use it in our experiments. Table 2 outlines the fairness metrics employed in the study. In Equation (7), NkNN represents the k-Nearest Neighbor function utilized to locate a specified number of instances (k = 5 in our scenario) surrounding xi within the attribute space. Ideally, these five neighbors should share the same label as xi. Any deviations from this expectation will result in a diminished consistency score, moving away from the perfect score of one.

Table 2: Fairness evaluation metrics

Notion	Formula	
Statistical Parity (SP)	$\left\lvert \frac{TP_p + FP_p}{N_p} - \frac{TP_u + FP_u}{N_u} \right\rvert$	(4)
Equal Opportunity (EO)	$\left\lvert \frac{TP_p}{TP_p + FN_p} - \frac{TP_u}{TP_u + FN_u} \right\rvert$	(5)
Equal Odds (EOdd)	$\frac{1}{2} * \left(\left\lvert \frac{FP_p}{FP_p + TN_p} - \frac{FP_u}{FP_u + TN_u} \right\rvert + \left\lvert \frac{TP_p}{TP_p + FN_p} - \frac{TP_u}{TP_u + FN_u} \right\rvert \right)$	(6)
Consistency N_{kNN}	$1 - \frac{1}{Nk} \sum_{i=1}^{N} \left\lvert y_i - \sum_{j \in N_{kNN}(x_i)} y_j \right\rvert$	(7)

We will compare our work with FLR [43, 48], "complement class" Naive Bayes (CNB) [49] for imbalanced dataset and original LR, NB models. We implement 2NB [43] and CNB [49] in Python within the scikit Learn framework, using Multinomial NB and download modified logistic regression FLR that applies fairness constraints to convex margin-based classifiers. In addition, we implement vanilla LR, Gaussian NB and our proposed method (3DBN). We then evaluate the models fairness and performance in 9 fairness benchmark datasets with different sensitive

Journal of Information Systems Engineering and Management
2025, 10(35s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

attributes (total 14 datasets) obtained from UCI repository [50], NYP [51] and [52].

In figure 3, we show the distribution of protected attribute values such as Female and Black compared to other attribute values such as Male and White. Each value presented with positive class (+ve) and negative class (-ve) outcome.

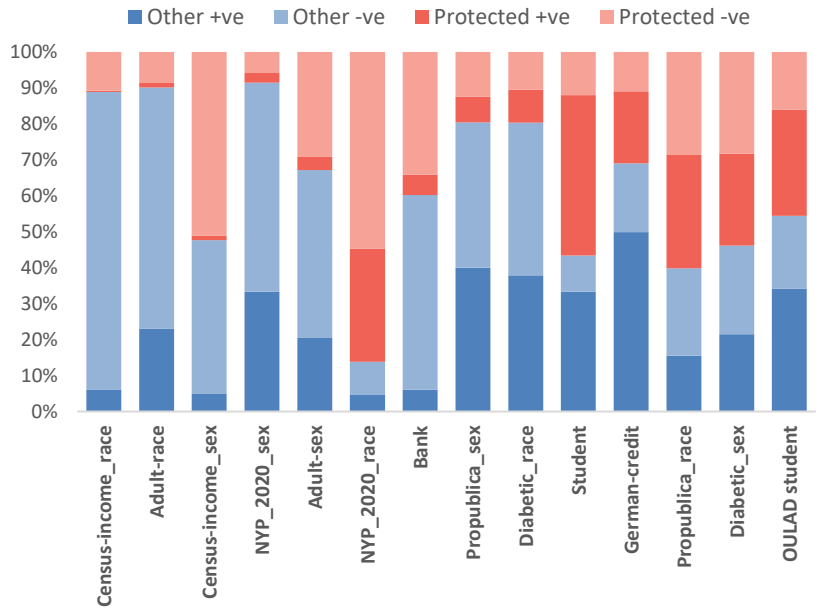


Figure 3: Protected attributes and class distribution

In Figure 4, we show the average performance (Accuracy and F-score), Individual fairness (Consistency), and Group fairness (SP, EOdd, and EO) for seven different algorithms on 14 datasets. The optimal result for (Acc, F1, and Consistency) is 1 and the optimal result for (SP, EOdd, and EO) is 0, however, the results is subtracted from 1 for consistency. Therefore, the optimal result is 100% for all metrics.

The result revels that our proposed algorithm (3DBN) consistently ranked in top 2 algorithms for all six metrics. FLR, LR and NB algorithms achieved optimal individual and group fairness for NYP 2020 and Census-income, respectively. However, the classifier's F1-score performance is 0. The FLR, LR and NB algorithms in these severely imbalanced datasets become a dummy classifier and predict all instances to be 0 (negative class). Our algorithm 3DBN, CNB, TAN, and 2NB classifiers don't suffer from this issue

Comparing to CNB, TAN, and 2NB, the proposed method (3DBN) significantly outperforming CNB for Accuracy, SP, EO, and EOdd. While CNB is significantly outperforming for consistently. Comparing to TAN and 2NB, our method 3DBN significantly outperforming for F1, EO, and EOdd. While 2NB is significantly outperforming for SP. It's worth mentioning that 2NB is optimized for SP and doesn't count for ground truth label and it focus on balanced outcome regardless of the causality between sensitive attribute and other attributes values. This could lead to Simpson's paradox as discussed earlier.

The finding revealed that not assessing multiple fairness metrics can lead to erroneous conclusions regarding the fairness performance of a model. It's possible for a model to be optimized for accuracy and individual/group fairness but still behave like a dummy classifier, predicting all instances as the majority class. In this scenario, the model attains high accuracy because it correctly predicts the majority class, and it achieves high individual/group fairness because there is no variation in outcomes for sensitive attribute values—they consistently predict the same outcome.

Moreover, focusing exclusively on optimizing group fairness based on sensitive attributes may lead to a distorted outcome. It could result in satisfactory group metric performance while compromising model accuracy, as it

Journal of Information Systems Engineering and Management
2025, 10(35s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

overlooks the inter-correlation between attributes. This situation is reminiscent of the Simpson's paradox observed in college admissions example, where the overall acceptance rate may seem unfair based on gender, but fairness is achieved when considering admission rates per gender within selected colleges.

Hence, the study recommends a thorough assessment of model fairness by concurrently examining multiple metrics. Additionally, it emphasizes cautious optimization, considering attribute inter-correlation and path-specific causal fairness, such as the influence of parent nodes on child nodes, as observed in the Bayesian network classifier.

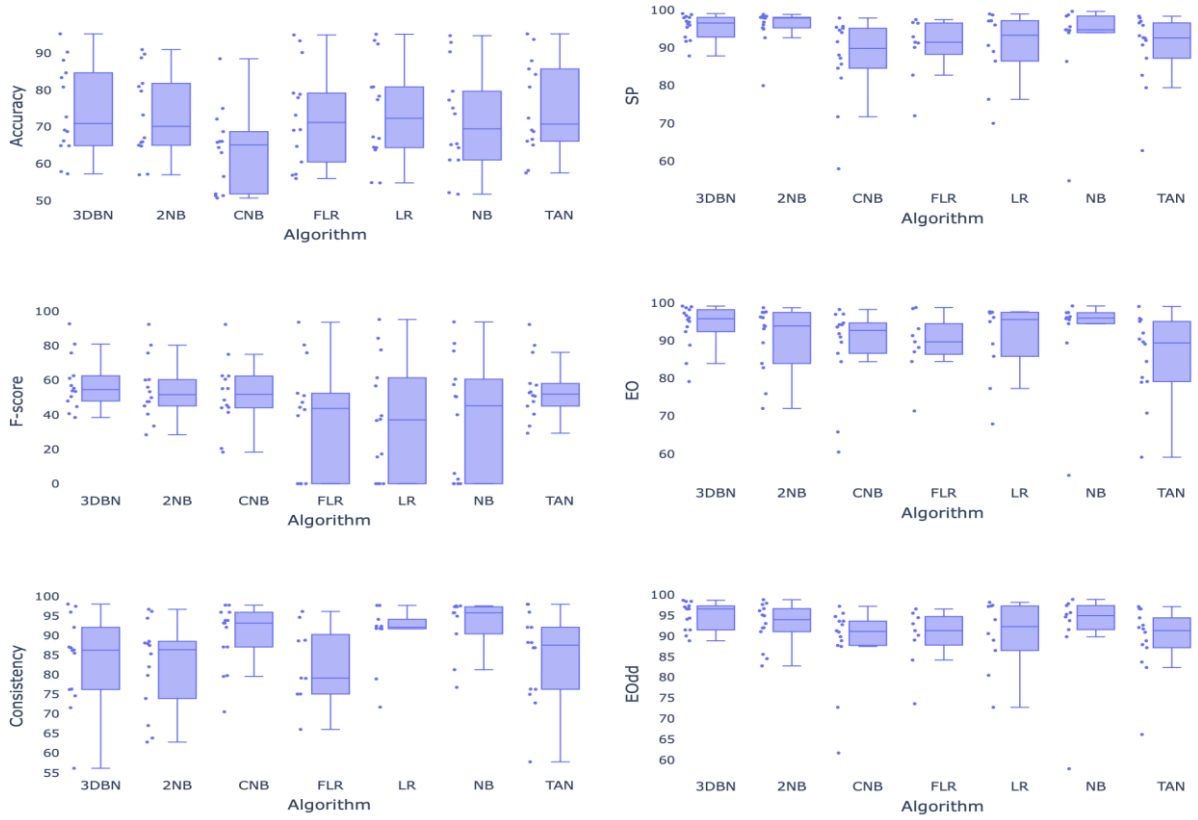


Figure 4: Classification performance and Fairness result for 3DNB compared with other classifiers

CONCLUSION

To sum up, we show that our method improves both fairness and F-score especially in real life dataset where the underlying representation for sensitive attribute often is bias due to scarce or very skewed data. This statistical and historical biases in the data will reflect the model bias against protected or minority groups. Therefore, in-process fine tuning mitigates the tradeoff between the model performance and fairness. Furthermore, our method dynamically finds implicit correlation between sensitive attribute value and other attribute values that could discriminate protected group and work on mitigate both terms probability and the underlying distribution.

In the presence of statistical biases, it is feasible to train a classifier that achieves high accuracy while giving an appearance of fairness based on statistical metrics such as Statistical party (SP) and Consistency (NkNN). However, the decision outcome might go completely injustice and predict all the outcome to one class. This is called dummy classifier, and metric such as Accuracy, SP, and Consistency will achieve high performance, while Fscore and EO, and EOdd only will detect this behavior. On the other hand, in the presence of historical bias, it also feasible to train a classifier to achieve 100% Statistical party in trade of with all other performance and fair metrics. However, in truth, the classifier may still be significantly unfair due to its causal dependence on sensitive attributes that cannot be justified without other attributes values causality. We see this example presented as Simpson's paradox.

The focus of our evaluation metrics is for two-classes problem, but the metric can be extended for multi-class

problem. Precisely, we are using Harmonic average as the sum of positive and negative classes, however, this will work for multi-class problem as well and we can use the harmonic average of the sum of all classes in (Eqs. 3). Similarly, evaluation metric can be extended to categorical and numeric sensitive attributes and not only binary attribute. The numeric values can be converted to nominal value using discretization, then, both categorical (ordinal or nominal) attributes will be evaluated by considering the fairness metrics in (Eqs. 4-7) between the most common value compared to each other attribute values. Fair Model should have values close to 0 for each pair of (common and minor) sensitive attributes values used in in (Eqs. 4-7). As future work, we intend to investigate using the complement harmonic average in different Bayesian Network classifiers and evaluate the result using different AI fairness metrics and benchmark datasets.

Acknowledgement:
This research was funded by Deanship of Scientific Research at King Saud University grant number RG-1439-035.

REFERENCES

[1] R. B. and K. A., "Extreme rebalancing for svms: a case study," *SIGKDD Explorations,* vol. 6, no. 1, pp. 60-69, 2004.

[2] G. &. C. E. Wu, "Class-Boundary Alignment for Imbalanced Dataset Learning.," *In ICML 2003 Workshop on Learning from Imbalanced Data Sets II, Washington, DC.,* 2003.

[3] R. Yan, Y. Liu, R. Jin and A. Hauptmann, "On predicting rare classes with SVM ensembles in scene classification," *IEEE International Conference on Acoustics, Speech and Signal Processing,* 2003.

[4] S. McGregor, Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database., In Proceedings of the Thirty-Third Annual Conference on Innovative Applications of Artificial Intelligence (IAAI-21). Virtual Conference., 2021.

[5] E. Commission, "High-level expert group on artificial intelligence (HLEG, AI). Ethics guidelines for trustworthy AI," *Tech. Rep., European Commission ,* 2019.

[6] S. Thiebes, S. Lins and A. Sunyaev, "Trustworthy artificial intelligence," *Electron. Markets ,* vol. 31, p. 447–464, 2021.

[7] A. Jobin, M. Ienca and E. Vayena, "The global landscape of AI ethics guidelines," *Nat. Mach. Intell.,* vol. 1, p. 389–399, 2019.

[8] S. Corbett Davies and S. Goel, "Defining and designing fair algorithms.," *https://policylab.stanford.edu/projects/defining-and- designing-fair-algorithms.html.,* 2018.

[9] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big data,* vol. 5, no. 2, p. 153–163, 2017.

[10] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel and A. and Huq, "Algorithmic decision making and the cost of fairness," *In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining ACM,* p. 797–806, 2017.

[11] M. Hardt, E. Price and N. Srebro, "Equality of opportunity in supervised learning.," *In Advances in neural information processing systems,,* p. 3315–3323., 2016.

[12] B. Green and L. Hu, "The myth in the methodology: towards a recontextualization of fairness in machine learning.," *ICML.,* 2018.

[13] Friedler, Scheidegger, Venkatasubramanian, Choudhary, Hamilton and Roth., "A comparative study of fairness-enhancing interventions in machine learning.," *In ACM Conference on Fairness, Accountability and Transparency (FAT*). ACM.,* 2019.

[14] A. Castelnovo, R. Crupi and G. e. a. Greco, "A clarification of the nuances in the fairness metrics landscape.," *Nature Sci Rep ,https://doi.org/10.1038/s41598-022-07939-1,* 2022.

[15] R. Binns, "On the Apparent Conflict Between Individual and Group Fairness," *Conference on Fairness, Accountability, and Transparency (FAT* '20), https://doi.org/10.48550/arXiv.1912.06883,* 2020.

[16] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman and A. Galstyan, "A survey on bias and fairness in machine learning.," *ACM Comput. Surv. (CSUR),* vol. 54, p. 1–35, 2021.

[17] C. Oneil, "Weapons of Math Destruction: How Big Data Increases Inequality and Threatens democracy," *Crown,London,* 2016.

[18] S. Barocas and A. Selbst, "Big data's disparate impact," *Calif. Rev.,* vol. 104, p. 671, 2016.

[19] J. Angwin, J. Larson, S. Mattu and L. Kirchner, "Machine bias: There's software used across the country to predict future criminals, and it's biased against blacks," *ProPublica,* 2016.

[20] D. Mulligan, J. Kroll, N. Kohli and R. Wong, "This thing called fairness: Disciplinary confusion realizing a value in technology," *Proc. ACM Hum. Comput. Interact. ,* vol. 3, pp. 1-36, 2019.

[21] S. Mitchell, E. Potash, S. Barocas, A. D'Amour and K. Lum, "Algorithmic fairness: Choices, assumptions, and definitions.," *Annu. Rev. Stat. Appl. ,* vol. 8, p. 141–163, 2021.

[22] A. Narayanan, "Fat* tutorial: 21 fairness definitions and their politics.," *New York, NY, USA.,* 2018.

[23] S. a. R. J. Verma, "Fairness definitions explained.," 2018.

[24] S. Mitchell, E. Potash and S. Barocas, "Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions.," *arXiv preprint arXiv:1811.07867.,* 2018.

[25] C. Dwork, M. Hardt, T. Pitassi, O. Reingold and R. and Zemel, "Fairness through awareness.," *In Proceedings of the 3rd innovations in theoretical computer science conference, ACM.,* p. 214–226, 2012.

[26] R. Nabi and I. Shpitser, "Fair inference on outcomes.," *In Proceedings of the AAAI Conference on Artificial Intelligence, ,* vol. 1931, 2018.

[27] C. Louizos, K. Swersky, Y. Li, M. Welling and R. and Zemel, "The variational fair autoencoder.," *arXiv preprint arXiv:1511.00830.,* 2015.

[28] G. Yona and G. and Rothblum, "Probably approximately metric-fair learning.," *In International Conference on Machine Learning,* p. 5666–5674, 2018.

[29] N. Kilbertus, M. R. Carulla, G. Parascandolo, M. Hardt, D. Janzing and B. and Scho ¨lkopf, "Avoiding discrimination through causal reasoning.," *In Advances in Neural Information Processing Systems,* p. 656–666., 2017.

[30] M. J. Kusner, J. Loftus, C. Russell and R. and Silva, "Counterfactual fairness.," *In Advances in Neural Information Processing Systems,,* p. 4066–4076., 2017.

[31] M. Buda, A. Maki and M. Mazurowski, "A Systematic Study of the Class Imbalance Problem in Convolutional Neural Networks," *Neural Networks ,* vol. 106, p. 249–259, 2018.

[32] H. H. and G. E.A., "Learning from Imbalanced Data," *Transactions on Knowledge & Data Engineering ,* vol. 9, p. 1263–1284, 2008.

[33] F. M., F. S. A., M. J., S. C. and S. and Venkatasubramanian, "Certifying and removing disparate impact.," *In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 259–268. ACM.,* 2015.

[34] S. Caton and C. Haas, FAIRNESS IN MACHINE LEARNING: A SURVEY, arXiv:2010.04053, 2019.

[35] S. Hajian and J. Domingo-Ferrer, "A methodology for direct and indirect discrimination prevention in data mining.," *IEEE transactions on knowledge and data engineering ,* vol. 25, no. 7, p. 1445–1459., 2013.

[36] F. Kamiran, I. Zliobaite and T. and Calders, "Quantifying explainable discrimination and removing illegal discrimination in automated decision making.," *Knowledge and information systems ,* vol. 35, no. 3, p. 613–644, 2013.

[37] I. Chen, F. D. Johansson and D. Sontag, "Why is my classifier discriminatory?," *Advances in Neural Information Processing Systems,* p. 3539–3550, 2018.

[38] B. T. Luong, S. Ruggieri and F. Turini, "kNN as an Implementation of Situation Testing for Discrimination Discovery and Prevention," p. 502–510, 2011.

Research Article

[39] R. Zemel, Y. Wu, K. Swersky, T. Pitassi and C. Dwork, "Learning Fair Representations," 2013.

[40] M. B. Zafar, I. V. Martinez, M. G. Rodriguez and K. P. Gummadi., Fairness Constraints: Mechanisms for Fair Classification, AISTATS, 2017.

[41] R. Berk, H. Heidari, S. Jabbari, M. Joseph, M. Kearns, J. Morgenstern, S. Neel and A. and Roth, "A convex framework for fair regression.," *arXiv preprint arXiv:1706.02409.,* 2017.

[42] S. G. S. Corbett Davies, "The measure and mismeasure of fairness: A critical review of fair machine learning.," *arXiv preprint arXiv:1808.00023.,* 2018.

[43] T. Calders and V. Sicco, "Three naive Bayes approaches for discrimination-free classification," *Data mining and knowledge discovery,* vol. 21, no. 2, pp. 277-292, 2010.

[44] S. Boulitsakis Logothetis, Fairness-aware Naive Bayes Classifier for Data with Multiple Sensitive Features, https://arxiv.org/abs/2202.11499, 2022.

[45] Z. Xianli and D. Edgar, "BAYES-OPTIMAL CLASSIFIERS UNDER GROUP FAIRNESS," *arXiv:2202.09724v3,* 2022.

[46] D. Paul and L. Michael, "Approximating probabilistic inference in Bayesian belief networks is NP-hard,," *Artificial Intelligence,* vol. 60, no. 1, pp. 141-153, 1993.

[47] F. S. Alenazi, K. El Hindi and B. AsSadhan, "Complement-Class Harmonized Naïve Bayes Classifier," *Applied Sciences ,* vol. 13, no. 8, p. 4852, 2023.

[48] M. B. Zafar, I. V. Manuel, G. Rodriguez and K. P. Gummadi, "Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment.," *In Proceedings of the 26th International Conference on World Wide Web (WWW '17),* 2017.

[49] D. M. Jason, L. Rennie, J. T. Shih and R. K. David, "Tackling the poor assumptions of naive bayes text classifiers.," *In Proceedings of the Twentieth International Conference on International Conference on Machine Learning (ICML'03),* 2003.

[50] D. Dua and C. Graff, "UCI Machine Learning Repository," *[http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.,* 2019.

[51] N. Y. P. D. NYPD, "Stop, Question, and Frisk Database," *Inter-university Consortium for Political and Social Research, 09 06 2008. [Online].,* p. Available: https://doi.org/10.3886/ICPSR21660.v1.

[52] J. Larson, S. Mattu, L. Kirchner and J. Angwin., "https://github.com/propublica/compas-analysis," 2016.