Journal of the Brazilian Computer Society, 2026, 32:1, doi: 10.5753/jbcs.2026.5960

Analysis of Performance and Fairness of Classification Algorithms with Class and Protected Attribute Noise

Ana Vargas [**National University of San Marcos** | ***anacecilia.vargas@unmsm.edu.pe***]
Rosa Delgadillo [**National University of San Marcos** | ***rdelgadilloa@unmsm.edu.pe***]

Postgraduate Program in System Engineering and Informatics, Universidad Nacional Mayor de San Marcos, Calle Germán Amézaga 375, Lima 15081, Perú.

Received: 25 April 2025 • **Accepted:** 11 May 2026 • **Published:** 22 June 2026

Abstract. Noise in data is an underexplored source of bias. This study investigates the impact of noise on both the performance and fairness of three classic classifiers: Random Forest (RF), Support Vector Machine (SVM), and Logistic Regression (RL), using four public datasets for binary classification. A comparative analysis is conducted to examine the effects of different types and levels of noise introduced in a predictor variable (protected attribute) and/or the response variable (class). Performance is measured using accuracy (ACC), while model fairness is evaluated using the Average Absolute Odds Difference (AAOD) metric, with a binary attribute as the protected. We results suggest that injecting noise into both the class label and the protected attribute (cn) yields results similar to injecting noise only into the labels (ln). The observed robustness of the Average Absolute Odds Difference (AAOD) under these conditions warrants careful interpretation. As AAOD values tend to decrease or remain stable despite increasing noise levels, there is a risk that the metric may be insensitive to noise. This suggests that AAOD might artificially report higher fairness in scenarios where significant underlying disparities actually persist.

Keywords: Classification, Noise model, Fairness, Accuracy

1 Introduction

Machine Learning (ML) models are rapidly spreading throughout society, increasingly influencing people's daily lives in various domains. They are used in critical situations where decisions have significant consequences on people's lives, such as decision-making in criminal justice, granting financial loans, selecting job candidates, allocating medical resources, and assessing insurance risks, among other applications.

However, these applications could rely on data that present limitations in data collection processes, such as measurement errors, incomplete data, distorted data, among others, meaning they exhibit noise to a greater or lesser extent. In classification models, noise can affect attributes or the class by altering their values and can result from inaccuracies in human-generated labels [Davani *et al*., 2022], errors in input attributes, imperfect algorithms for inferring protected attributes [Ghosh *et al*., 2021], or intentional data tampering [Ghosh *et al*., 2022]. The learning process with noisy data can hinder this learning and bias the results, making classifiers less accurate in their performance.

On the other hand, noise in data is an underexplored source of bias, according to [Corbett-Davies *et al*., 2017], which can exacerbate existing social discrimination against minority or vulnerable groups, such as women, indigenous peoples, migrants, etc., leading to a detriment in social welfare and cohesion [Barocas *et al*., 2017]. Therefore, several researchers have focused on developing classifiers that jointly optimize predictive performance and fairness [Ghosh *et al*., 2023]. [Caton and Haas, 2024] compiles and classifies the proposals of several of these algorithms that incorporate one or more fairness metrics into the optimization functions of machine learning algorithms with the objective of maximizing performance and fairness, among the most recent of which are [Yeom and Fredrikson, 2020], [Gao *et al*., 2022], [Grazzi *et al*., 2022], [Hossain *et al*., 2021], [Huang *et al*., 2022], [Stefano *et al*., 2020], [Blanzeisky and Cunningham, 2022], [Ferry *et al*., 2023], [Roh *et al*., 2021] and [Patil *et al*., 2021].

Since noise in real-world data is not quantifiable and its characteristics are unknown, different studies have introduced synthetic noise experimentally and in a controlled manner, allowing conclusions to be drawn according to the type of noise, its frequency, and its characteristics. Most of these studies have analyzed the effect of noise on the label [Canalli *et al*., 2024], [Silva *et al*., 2024], [Wang *et al*., 2021], [Wu *et al*., 2022] and [Xia *et al*., 2020] or the effect of noise on the protected attribute [Ghosh *et al*., 2023], [Celis *et al*., 2021] and [Khani and Liang, 2020] and very few in both conditions.

To deepen the understanding of the relationship between noise in data—whether in the classes, attributes, or both—and its impact on the performance and fairness of classification models, this study aims to explore empirically how noise can affect the performance metric and fairness metric in binary classifiers. A comparative analysis is conducted to examine the effect of incorporating different types and levels of noise into a predictor variable (attribute) and/or the response variable (class) in three traditional classification algorithms, more used as baseline models, using four public datasets.

Comparative evaluations, such as the one presented in this study, are essential for researchers and machine learning practitioners, as the findings can serve as a reference by providing the scientific community with valuable information about the relative performance of these classical classifiers. Moreover,

they can offer guidelines for selecting the "best" classifiers in real-world scenarios and highlight areas that still require the development of new algorithms, providing a solid framework for rigorous evaluation.

This study is structured as follows: Section 2 offers a brief introduction to related work, baseline classification algorithms, and the metrics used to assess predictive performance and fairness. Section 3 details different techniques employed to introduce noise into the dataset. Section 4 outlines the experimental approach, including a description of the datasets used. Finally, Section 5 presents the results obtained and Section 6 the conclusions derived from the study.

2 Related work

In the context of classification models, studies have focused on the effect of class noise, paying little attention to the impact of attribute noise in datasets and scant attention to the combined impact in both situations [Sáez, 2022]. According to the systematic review by Gupta and [Gupta and Gupta, 2019], class noise impacts more than attribute noise, but both degrade performance.

Within the context of fair classification models, [Angwin *et al*., 2016] initiated debates about biases in algorithms within systems that automate decisions. [Dwork *et al*., 2012] introduced the concept of fairness, and [Zemel *et al*., 2013] proposed statistical parity and fair learning methods. Arising from these are various proposals for metrics to measure fairness and mitigation algorithms [Hardt *et al*., 2016]; [Chouldechova, 2017];[Mehrabi *et al*., 2021]. Comparative studies of different algorithms are standard practice among machine learning researchers [Ghosh *et al*., 2023]. Among the comparative studies of fair classification algorithms are those by [Friedler *et al*., 2019], who show that there is a trade-off between fairness and accuracy in many of these algorithms, and those by [Hort *et al*., 2021], who provide a conceptual framework for the balance between them.

All these comparative studies start from baseline models (reference models without any fairness constraints), such as Logistic Regression (LR), Random Forest (RF), Decision Trees, Support Vector Machines (SVM), Naive Bayes, among others. However, these studies have not delved into the differences among them in the presence of noise, leaving questions about how the impact of noise in data varies in fairness metrics and the trade-off between fairness and accuracy unanswered. It also raises the question of whether this impact depends on the type of baseline model chosen.

2.1 Classifiers

The specialized literature has introduced various classification methods. According to the survey conducted by Hort *et al*. [2024], the algorithms Random Forest, Logistic Regression, Neural Network and Support Vector Machine are among the top four most frequently used algorithms.

2.2 Model Performance and Fairness Metrics

Currently, it is common to evaluate the accuracy of a model's predictions and its performance using loss functions. These metrics compare the actual or known value with the prediction made by the model. The goal of these functions, also known as error functions, is to minimize them, which results in a more accurate model. There are various types of loss functions, depending on whether the problem is regression or classification [Khani, 2021]. The most widespread and commonly used method for evaluating the performance of a classification model is accuracy. This measure represents the proportion of correct predictions made by the model compared to the total number of predictions. Most studies that assess the effect of noise on classification models evaluate performance and accuracy using this metric [Alharbi, 2024]; [Hasan and Chu, 2022].

On the other hand, fairness refers to the development of models that do not introduce or perpetuate unfair biases against specific groups. Ensuring fairness is crucial to making sure that automated decisions are just and non-discriminatory, especially in sensitive applications such as hiring, loan approvals in the financial system, or legal decisions in the judicial system. To measure fairness in classification models, the literature has proposed metrics through two main approaches [Mehrabi *et al*., 2021]; [Pessach and Shmueli, 2022]; [Caton and Haas, 2024]:

- **Group fairness (statistical)**: This approach is based on assessing fairness through statistics and metrics calculated for different demographic groups or subpopulations within the overall population. These groups are typically defined based on sensitive attributes such as gender, race, age, etc. The goal is to ensure that the model does not discriminate against any of these groups.

- **Individual fairness (counterfactual)**: This approach focuses on fairness at the individual level by comparing the model's decisions for an individual with the decisions it would have made for the same individual in a counterfactual world where certain sensitive attributes are altered. The principle is that the model's outcome should not change significantly if an individual's sensitive attributes were altered.

The group-based approach has used as the most prominent framework in the literature [Mehrabi *et al*., 2021]. Its associated metrics are essential for evaluating and ensuring that classification models do not exhibit bias against specific sensitive or protected groups.

One of these metrics is the Average Odds Difference ($AAOD$). It is based on the confusion matrix and combines two aspects related to equal odds (EO): false positives (FP) and false negatives (FN), which are common errors in classification models. This metric is defined as the average of the absolute difference in the False Positive Rate (FPR) and the True Positive Rate (TPR) between the privileged or protected group (A) and the unprivileged or unprotected group (B). This metric measures the average of the absolute probability differences between different demographic groups for FN and FP.

[Ghosh *et al*., 2023] use a similar metric as a tool to assess the fairness of the model between different demographic groups. The AAOD measure is defined as follows:

$$AAOD = \frac{1}{2}|FPR_A - FPR_B| + \frac{1}{2}|TPR_A - TPR_B| \quad (1)$$

where:
FPR: false positive rate for A or B
TPR: True positive rate for A or B

An $AAOD$ of 0 indicates that the model is perfectly fair, as there are no differences in false positive and false negative rates between the groups. An $AAOD$ greater than 0 indicates that there are differences in error rates between groups, suggesting bias in the model.

To clarify the terminology used in this study, consider a classification task on the Adult Census dataset, where the goal is to predict whether an individual has income level more than \$50000 (high-income). Here, 'sex' serves as the sensitive attribute. This attribute defines two distinct protected groups: 'Male' (often the privileged group) and 'Female' (the unprivileged group). When we refer to Group Fairness, we evaluate whether the model's predictions—such as the probability of being classified as 'high-income' —are balanced across these demographic segments, regardless of their underlying distribution.

3 Noise model in data

[Sáez, 2022] indicates that many classification algorithms are designed to handle data in which noise may exhibit particular characteristics. However, the noise affecting real-world data is often difficult to quantify, and its characteristics are largely unknown. To assess the effectiveness of these methods in a controlled environment, noise models have been proposed that introduce errors into datasets in a supervised manner. These models facilitate control over the type, amount, and characteristics of the added noise, allowing the design of experimental frameworks suitable for the objective of the study and the extraction of meaningful conclusions. Their use enables, for instance, the study of circumstances under which classifiers are most affected (depending on the type and level of noise) or the evaluation of the effectiveness of noise pre processing techniques to detect and correct introduced errors. [Sáez, 2022] structures the elements and configuration of noise models into three key components: the selection procedure, the type of noise, and the alteration procedure.

The selection procedure involves identifying the values (indexes) of the attribute and/or tags that will be altered. The noise type refers to the variable(s) affected by the noise model, which may involve: (a) the class, (b) the attribute, or (c) both, i.e., both the class and the attribute. Finally, the alteration procedure defines how the selected values will be modified during the selection process.

4 Experimental analysis

Four distinct datasets, all focused on binary classification tasks, were selected following the review results given by [Hort *et al*., 2024], we used three of the most established datasets in the literature: Adult, COMPAS, and German. The fourth dataset, OULAD, was included as it is one of the most recent, as noted by [Le Quy *et al*., 2022].

To assess fairness, a binary sensitive or protected attribute—such as gender or race—was considered to determine whether the models are fair toward both groups defined by this attribute.

1. **OULAD** [Kuzilek and Zdrahal, 2015]: The dataset contains information of students and their activities in the virtual learning environment (VLE) for seven courses. Following data cleaning and exclusion of missing values, the final sample consisted of 31 482 samples and 8 attributes. The prediction task was on the class label 'final_result' (pass, non_pass) and 'gender' was used as the protected variable for the analysis as shown in **Table 1**.
2. **Adult** [Becker and Kohavi, 1996]: Its goal is to predict whether annual income of an individual exceeds \$50K/yr based on census data. Following data cleaning and exclusion of missing values, the final sample consisted of 48842 and 12 attributes. The target variable is income, indicating whether an individual makes less or more than 50K, and 'sex' (male or female) was used as the protected attribute. An overview of attribute characteristics is shown in **Table 2**.
3. **Compas** [Angwin *et al*., 2016]: It contains information used by the COMPAS system, a proprietary software developed by Northpointe, which is used in US criminal justice systems to predict the risk of criminal recidivism. After cleaning and filtering out missing values, the dataset was reduced to 6 172 instances and 8 features whose description is showed in **Table 3**. The 'two_year_recid' variable was designated as the target, and race (Caucasian/Other) was the protected attribute.
4. **German** [Hofmann, 1994]: The dataset includes 1000 complete instances (with no missing values), with the response class serving as the target variable. Initially, the initial 'Per-stat' attribute was decomposed to dissociate sex/gender from personal status, resulting in two separate attributes: 'marital status' and 'sex'. The sex attribute was designated as the protected variable for the analysis. The attributes description is shown in **Table 4**.

An experiment was designed to evaluate the performance and fairness of three classical classifiers: Random Forest, Support Vector Machine, and Logistic Regression, in the four datasets. In addition, this experiment aims to offer a deeper understanding of the theoretical fairness and predictability limits of these classifiers when the data contains noise.

According to the nomenclature proposed by [Sáez, 2022], the noise model utilized is the *symmetric uniform label (attribute or both)*, where the procedure for selecting the cases to be altered is random with equal probability, and the disruption or alteration procedure consists of flipping the values of

Table 1. OULAD dataset. Attributes characteristics

Atribute	Type	Values	Description
gender	Categorical	2	**Protected attribute**. Gender
region	Categorical	13	The geographic region
highest_education	Categorical	5	The category of the highest student education level
imd_band	Categorical	10	The index of multiple deprivation (IMD) band of the place where the student lived
age_band	Categorical	3	The category of the student's age
num_of_prev_attempts	Numerical	0 - 6	The number times the student has attempted this module
studied_credits	Numerical	30 - 655	The total number of credits for the modules the student is currently studying
disability	Categorical	2	Whether the student has declared a disability
final_result	Categorical	2	**Target class**. Whether the student's final result was pass or non with a distribution ratio of 53:47, respectively.

Table 2. Adult dataset. Attributes characteristics

Attribute	Type	Values	Description
age	Numerical	17 - 90	The age of an individual
workclass	Categorical	4	The employment status
education-num	Numerical	1 - 16	The highest level of education achieved in numerical form
marital-status	Categorical	4	The marital status.
occupation	Categorical	3	The general type of occupation
relationship	Categorical	5	Represents what this individual is relative to others
race	Categorical	2	Race
sex	Categorical	2	**Protected attribute**. The biological sex of the individual.
capital-gain	Numerical	0 - 99999	The capital gains for an individual
capital-loss	Numerical	0 - 4356	The capital loss for an individual
hours-per-week	Numerical	1 - 99	The hours an individual has reported to work per week
native-country	Categorical	2	Whether or not an individual's country of origin is the US
income	Categorical	2	**Target class**. Whether or not an individual makes more than $50,000 annually with a distribution ratio of 25:75, respectively.

Table 3. COMPAS dataset. Attributes characteristics

Attribute	Type	Values	Description
c_charge_degree	Categorical	2	The degree of the current charge (F: Felony, M: Misdemeanor)
race	Categorical	2	**Protected attribute**. The race/ethnicity of the defendant
age_cat	Categorical	3	Age category of the defendant
sex	Categorical	2	The sex of defendent
priors_count	Numerical	0 - 38	The number of prior criminal offenses/convictions
two_year_recid	Categorical	2	**Target class**. Whether the defendant was rearrested within two years or non with a distribution ratio of 46:54, respectively.
length_of_stay	Numerical	-1 - 799	Calculated number of days the defendant stayed in jail.
juv_crime	Numerical	0 - 20	Calculation of the total number of prior offenses committed by an individual while they were a minor.

Table 4. German dataset. Attributes characteristics

Attribute	Type	Values	Description
StaAcc	Categorical	4	The status of existing checking account
DuMon	Numerical	4 - 72	Duration of the credit (month)
CredHis	Categorical	5	The credit history
Purpose	Categorical	10	Purpose (car, furniture, education, etc.)
CredAmt	Numerical	250 - 18424	Credit amount
SaveAcc	Categorical	5	Saving account/bonds
PreEmpl	Categorical	5	Present employment since.
InsRt	Numerical	1 - 4	The installment rate in percentage of disposable income
OthDebtor	Categorical	3	Other debtors / guarantor.
PreRe	Numerical	1 - 4	Present residence since
Property	Categorical	4	Property
Age	Numerical	19 - 75	The age in years
IntPla	Categorical	3	Other installment plans
Housing	Categorical	3	Housing (rent, own, for free)
ExstCredit	Numerical	1 - 4	Number of existing credits at this bank
Job	Categorical	4	Job
NoMain	Numerical	1 - 2	Number of people being liable to provide maintenance for
Phone	Categorical	2	Telephone
ForWorker	Categorical	2	foreign worker
Response	Categorical	2	**Target class**. The customer's level of risk (good or bad) with a distribution ratio of 7:3 respectively.
sex	Categorical	2	**Protected attribute**. The sex
marital	Categorical	2	The marital status

the label or the protected attribute, as illustrated in **Figure 1**. Noise was introduced in three different ways, depending on the type of noise:

- **Noise in the class (ln)**: Noise is introduced into class labels with 5-level: 0%, 10%, 20%, 30% or 40% of the label class's values flipped.
- **Noise in protected attribute (an)**: Noise is introduced into protected attribute with 5-level: 0%, 10%, 20%, 30% or 40% of the protected attribute's values flipped.
- **Combined noise (cn)**: Noise is introduced to both the class and the protected attribute. It should be noted that the procedure for selecting cases to be altered was performed twice, once to alter the labels and once to alter the attribute. This implies that there could exist cases where the values of both the labels and the attribute were flipped, and consequently, the percentage of these cases was recorded in each repetition.

After adding the noise, the data were randomly divided into a 80:20 ratio for the training and test sets. The split is performed by ensuring that the training and test sets have approximately the same proportion of observations from each target class as the initial data set. This is essential to avoid bias, especially with unbalanced datasets. The data are modeled using one of the algorithms under study, and the accuracy and fairness ($AAOD$) metrics are calculated from the predictions.The experiment is repeated 10 times for each noise type and level, using different data partitions for the training and test sets to ensure statistical reliability and evaluate the stability of each classifier's metrics.

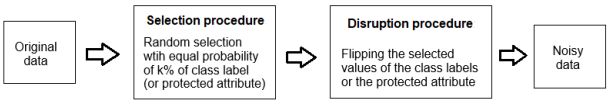


Figure 1. Experimental process for altering the dataset

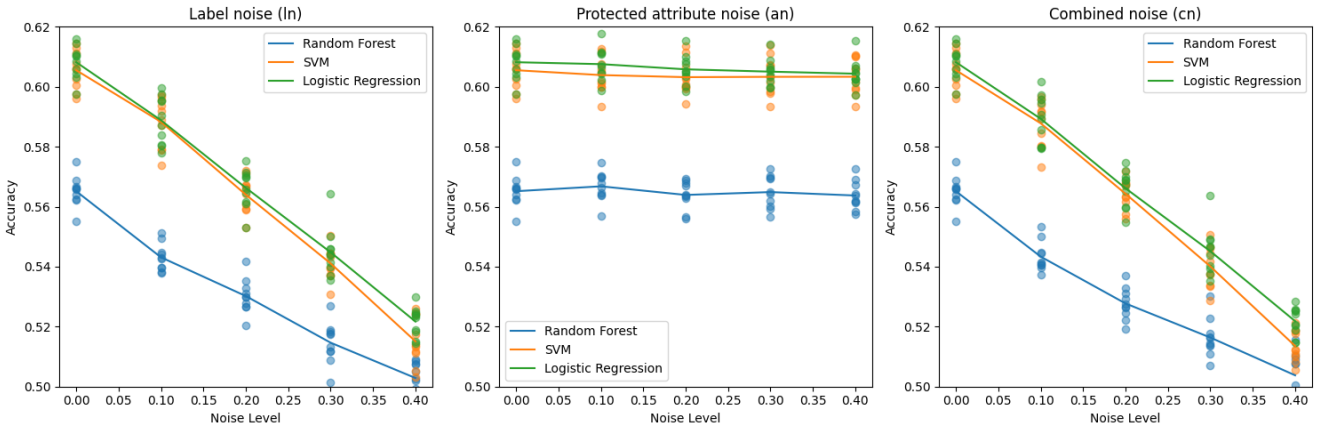
5 Results and discussion

In this section, the results of the experiments of the three classification methods are described, showing their performance in the presence of synthetic noise. Figures 2–5 show the measures of accuracy and fairness that are taken along the levels of added noise for each classifier in the four datasets. The solid lines correspond to the average of the ten runs for each noise level.

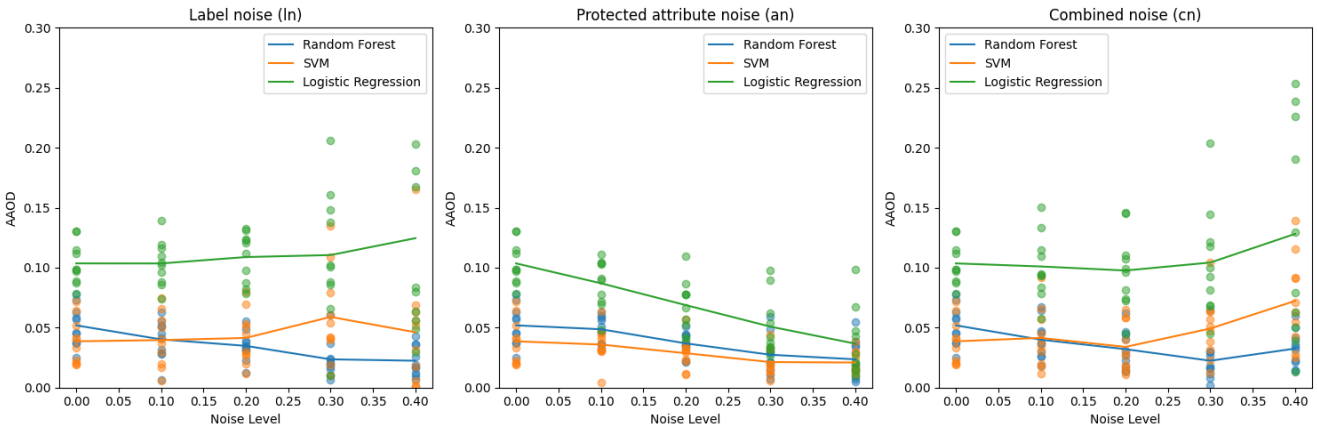
For the OULAD dataset (as shown in Figure 2), accuracy is observed to decline as the noise level increases when noise is added solely to the label (ln) or when it is added to both the label and the protected attribute (cn). This trend, however, is not observed when noise is added only to the protected attribute (an). For the classifiers, the graph displays the Random Forest has lower accuracy values compared to the other two methods.

In the same figure, the averages of the $AAOD$ equity measurements appear more stable when noise is added to the label (ln) or to both (cn). However, the behavior appears to slightly diverge when noise is added to the protected attributes (an). Slightly higher values are observed with Logistic Regression than with the other two methods.

Considering the results for the Adult dataset, shown in Figure 3, the ACC values are practically decreased by the increase in noise when noise is added solely to the labels (ln) or to both the labels and the protected attributes (cn), com-



(a) ACC over OULAD



(b) AAOD over OULAD

Figure 2. Accuracy and AAOD for OULAD dataset. Each point is the average of ten runs for the given classifier, dataset and noise level.

pared to the values that tend to remain stable when the noise is added only to the protected attributes (an). The $AAOD$ values exhibit an apparent decrease when noise is applied solely to the protected attribute (an). Conversely, this behavior is not replicated when noise is added to the label (ln) or to both features (cn)

Taking into account the results for the German dataset (shown in Figure 4), it is observed that the ACC measurements exhibit quite similar behavior across the three classifiers. Clearly, these values decrease when subjected to label noise (ln) or combined noise (cn), but they do not decrease when the noise type is attribute noise (an). Regarding the equity measurement ($AAOD$), a certain stability is also observed across all three types of noise (label noise, attribute noise, and combined noise: ln, an, and cn).

For the Compas dataset, the ACC values do not exhibit a complete tendency to decrease when noise is applied to the label (ln) or combined (cn), unlike the other datasets. Furthermore, when attribute noise (an) is introduced, ACC shows no significant change, similar to the observations in the other datasets.

It is observed that for the four datasets, the ACC measure exhibit similar behavior across all three algorithms and at all noise levels when noise is added to the labels (ln) or when noise is added to labels and protected attribute (cn). It was observed that the accuracy decreases as the noise level increases, except in the Compas dataset. However, when noise is added to the protected attribute, the accuracy level is observed to remain relatively constant across all noise levels. On the other hand, it is observed that the Random Forest classifier exhibits lower performance than the other algorithms, except in the German dataset.

Regarding the observed $AAOD$ values, a similar pattern in behavior is observed when noise is injected into the label (ln) compared to when noise is injected into both (cn) in almost all datasets, with slight differences in the Adult dataset. When noise is added to the protected attribute, the figures show that the $AAOD$ values tend to decrease slightly as the noise level increases, except for the German dataset. Furthermore, for this dataset, the behavior of these values remains stable across all noise levels for any noise type (ln, an, and cn) and even for any of the classifiers.

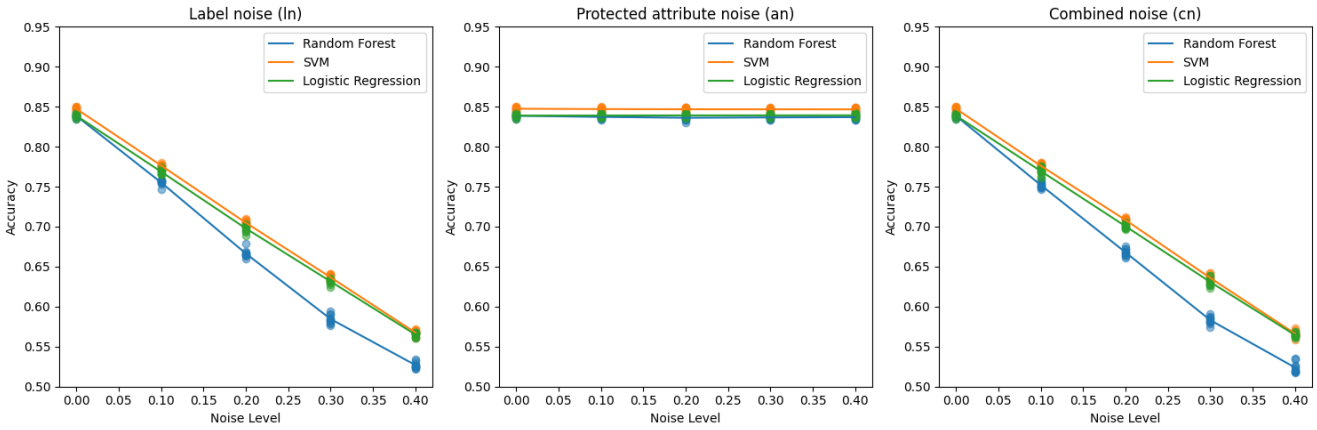
6 Conclusions and recommendations

In general, the effect of noise does not appear to be significant when introduced into the protected attribute (an), as the accuracy metric demonstrates strong robustness, and the equity metric also shows robustness, albeit to a lesser extent, as the percentage of noise in the data increases. Conversely, this phenomenon does not occur when noise is injected solely into the labels (ln) or into both attributes and labels (cn), as a decrease in model performance is observed. Therefore, if a dataset is suspected to contain noise primarily in the class label, it can be assumed that the accuracy will be significantly affected as the noise level increases. This is not the case for the fairness metric, which appears to be more robust.

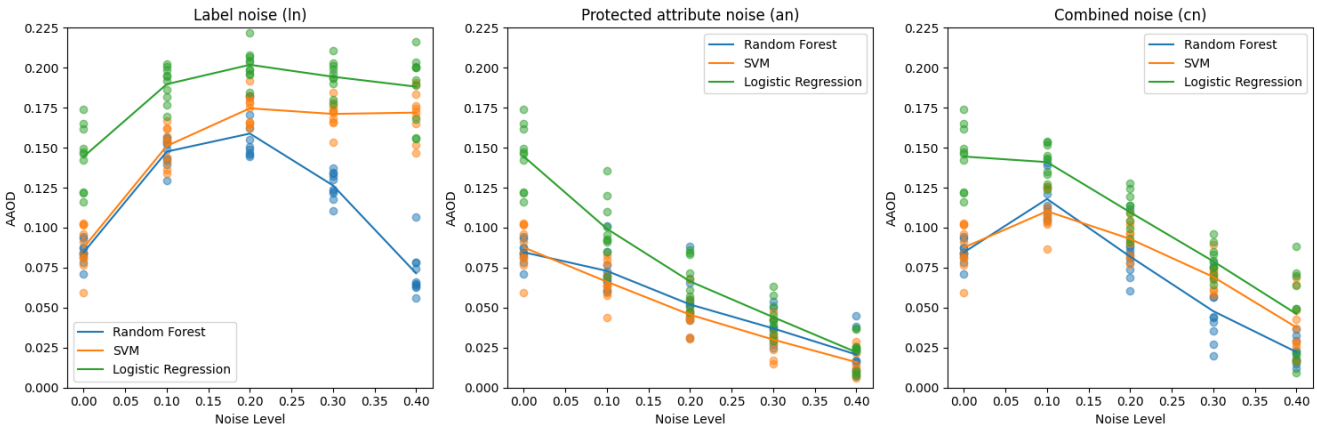
A point to highlight is the similarity in the behavioral trend of the values for both metrics (ACC and $AAOD$) across the different noise levels between ln and cn. This suggests that injecting noise into both the class label and the protected attribute (cn) yields results similar to injecting noise only into the labels (ln). Furthermore, it should be noted that this behavior differs when noise is injected solely into the protected attribute (an), specifically, the ACC metric does not appear to be affected as the noise level increases, while the $AAOD$ metric shows an apparent slight decrease as the noise level rises.

Regarding the fairness metric, the impact of noise tends to decrease or at least remain constant as the noise level is increased. This suggests that the metric is either insensitive to noise levels or that the observed trend is driven by the noise itself: as randomness obscures the underlying data structure across all groups, inter-group disparities diminish, leading to a convergence toward fairness. Consequently, higher noise levels may artificially inflate fairness scores, raising the concern that the metric may report equity where none actually exists.

This study presents some limitations, including the exclusive evaluation of binary classifiers, the use of a single equity metric based on a group-based approach, the consideration of only one binary sensitive attribute, and, from an experimental perspective, the simultaneous injection of noise only into protected attribute of the dataset. Therefore, a deeper exploration of the effect of noise on other inequity metrics constructed under the same or different approaches, as well as studying the effect considering a non-binary sensitive attribute or multiple sensitive attributes, or conducting more controlled experimentation by adding noise only to the sensitive attribute and analyzing its effect, could be of interest to the fair machine learning research community.

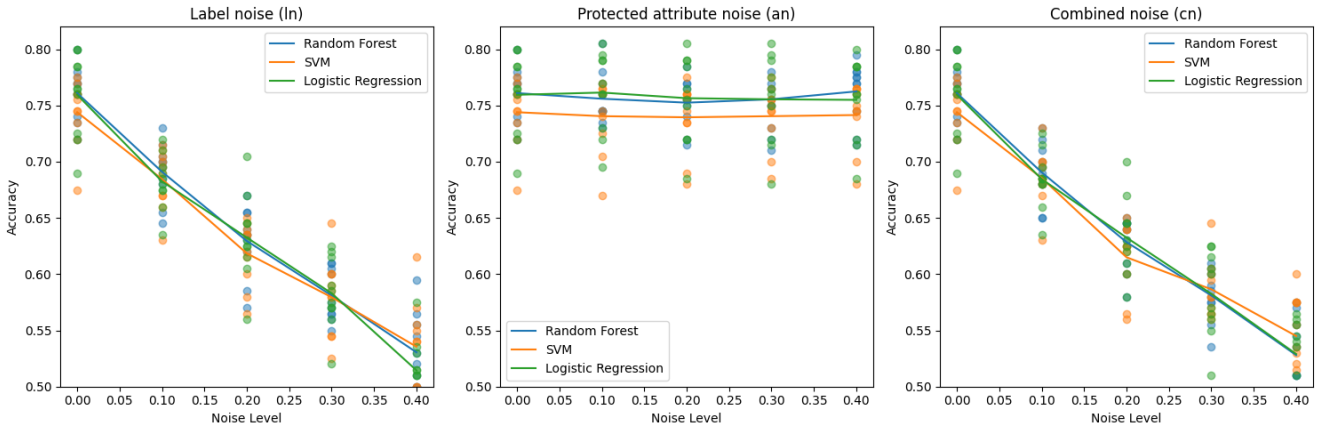


(a) ACC over Adult

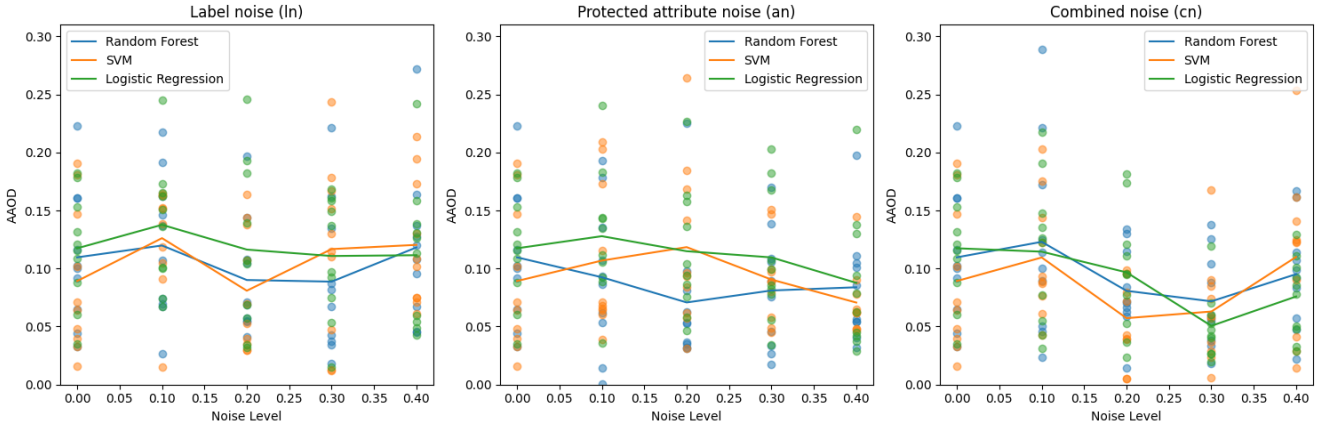


(b) AAOD over Adult

Figure 3. Accuracy and AAOD for Adult dataset. Each point is the average of ten runs for the given classifier, dataset and noise level.

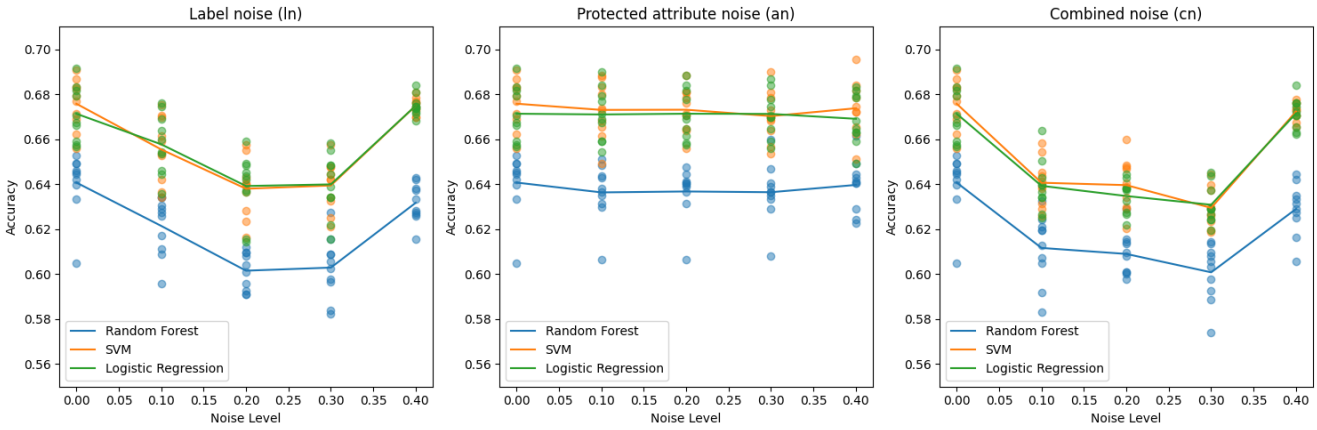


(a) ACC over German

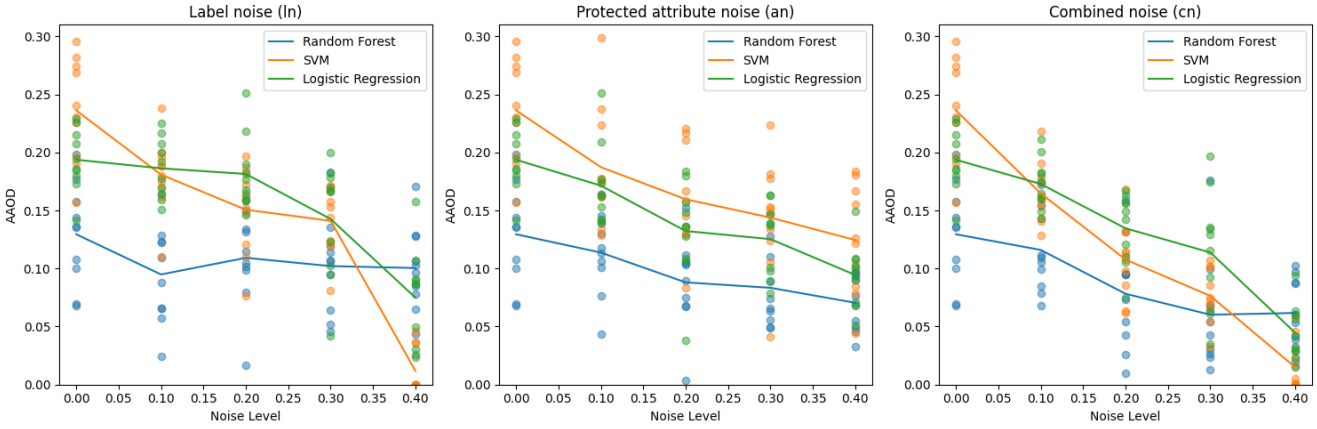


(b) AAOD over German

Figure 4. Accuracy and AAOD for German dataset. Each point is the average of ten runs for the given classifier, dataset and noise level.



(a) ACC over Compas



(b) AAOD over Compas

Figure 5. Accuracy and AAOD for Compas dataset. Each point is the average of ten runs for the given classifier, dataset and noise level.

Declarations

Authors' Contributions

AV: Contributed to the conception of the study, conceptualization of the methodology, literature review, algorithm selection, construction of software, data curation, formal analysis, writing—original and editing.
RD: Contributed to the conception of the study, conceptualization of the methodology, formal analysis, validation, writing—review and editing All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Funding

This research was not funded.

Availability of data and materials

Four datasets used in the work are public: Adult [Becker and Kohavi, 1996], Compas [Angwin *et al*., 2016], German [Hofmann, 1994] and OULAD [Kuzilek and Zdrahal, 2015] Source code: `https://github.com/unalmdei/noise`

References

Alharbi, A. A. (2024). Classification performance analysis of decision tree-based algorithms with noisy class variable. *Discrete Dynamics in Nature and Society*, 2024(1):6671395. DOI: 10.1155/2024/6671395.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals and it's biased against blacks. *ProPublica*. Available at:`https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`.

Barocas, S., Hardt, M., and Narayanan, A. (2017). Fairness in machine learning. NIPS Tutorial. Available at: `https://fairmlclass.github.io/`.

Becker, B. and Kohavi, R. (1996). Adult. UCI Machine Learning Repository. DOI: 10.24432/C5XW20.

Blanzeisky, W. and Cunningham, P. (2022). Using pareto simulated annealing to address algorithmic bias in machine learning. *The Knowledge Engineering Review*, 37:e5. DOI: 10.1017/S0269888922000029.

Canalli, Y., Braida, F., Alvim, L., and Zimbrão, G. (2024). Fair transition loss: From label noise robustness to bias mitigation. *Knowledge-Based Systems*, 294:111711. DOI: 10.1016/j.knosys.2024.111711.

Caton, S. and Haas, C. (2024). Fairness in machine learning: A survey. *ACM Computing Surveys*, 56(7):1–38. DOI: 10.1145/3616865.

Celis, L. E., Huang, L., Keswani, V., and Vishnoi, N. K. (2021). Fair classification with noisy protected attributes: A framework with provable guarantees. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1349–1361. PMLR. Available at:`https://proceedings.mlr.press/v139/celis21a.html`.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163. DOI: 10.1089/big.2016.0047.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pages 797–806. DOI: 10.1145/3097983.3098095.

Davani, A. M., Díaz, M., and Prabhakaran, V. (2022). Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110. DOI: 10.1162/tacl_a_00449.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226. DOI: 10.1145/2090236.2090255.

Ferry, J., Aivodji, U., Gambs, S., Huguet, M.-J., and Siala, M. (2023). Improving fairness generalization through a sample-robust optimization method. *Machine Learning*, 112(6):2131–2192. DOI: 10.1007/s10994-022-06191-y.

Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., and Roth, D. (2019). A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 329–338. DOI: 10.1145/3287560.3287589.

Gao, X., Zhai, J., Ma, S., Shen, C., Chen, Y., and Wang, Q. (2022). Fairneuron: improving deep neural network fairness with adversary games on selective neurons. In *Proceedings of the 44th International Conference on Software Engineering*, ICSE '22, page 921–933, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3510003.3510087.

Ghosh, A., Dutt, R., and Wilson, C. (2021). When fair ranking meets uncertain inference. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 1033–1043. DOI: 10.1145/3404835.3462850.

Ghosh, A., Jagielski, M., and Wilson, C. (2022). Subverting fair image search with generative adversarial perturbations. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 637–650. DOI: 10.1145/3531146.3533128.

Ghosh, A., Kvitca, P., and Wilson, C. (2023). When fair classification meets noisy protected attributes. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 679–690. DOI: 10.1145/3600211.3604707.

Grazzi, R., Akhavan, A., Falk, J. I., Cella, L., and Pontil, M. (2022). Group meritocratic fairness in linear contextual bandits. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A., editors, *Advances in Neural Information Processing Systems*, volume 35, pages 24392–24404. Curran Associates, Inc.. DOI: 10.52202/068431-1771.

Gupta, S. and Gupta, A. (2019). Dealing with noise problem in machine learning data-sets: A systematic review. *Procedia Computer Science*, 161:466–474. DOI: 10.1016/j.procs.2019.11.146.

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29. DOI: 10.48550/arXiv.1610.02413.

Hasan, R. and Chu, C. (2022). Noise in datasets: What are the impacts on classification performance?[noise in datasets: What are the impacts on classification performance?]. In *Proceedings of the 11th International Conference on Pattern Recognition Applications and Methods*. DOI: 10.5220/0010782200003122.

Hofmann, H. (1994). Statlog (German Credit Data). UCI Machine Learning Repository. DOI: 10.24432/C5NC77.

Hort, M., Chen, Z., Zhang, J. M., Harman, M., and Sarro, F. (2024). Bias mitigation for machine learning classifiers: A comprehensive survey. *ACM Journal on Responsible Computing*, 1(2):1–52. DOI: 10.1145/3631326.

Hort, M., Zhang, J. M., Sarro, F., and Harman, M. (2021). Fairea: A model behaviour mutation approach to benchmarking bias mitigation methods. In *Proceedings of the 29th ACM joint meeting on European software engineering conference and symposium on the foundations of software engineering*, pages 994–1006. DOI: 10.1145/3468264.3468565.

Hossain, S., Micha, E., and Shah, N. (2021). Fair algorithms for multi-agent multi-armed bandits. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 24005–24017. Curran Associates, Inc.. DOI: 10.48550/arXiv.2007.06699.

Huang, W., Labille, K., Wu, X., Lee, D., and Heffernan, N. (2022). Achieving user-side fairness in contextual bandits. *Human-Centric Intelligent Systems*, 2(3):81–94. DOI: 10.1007/s44230-022-00008-w.

Khani, F. (2021). *Causes, Measurement, and Mitigation of Loss Discrepancy*. Stanford University. Available at:`https://purl.stanford.edu/gw991vt5365`.

Khani, F. and Liang, P. (2020). Feature noise induces loss discrepancy across groups. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5209–5219. PMLR. Available at:`https://proceedings.mlr.press/v119/khani20a.html`.

Kuzilek, Jakub, H. M. and Zdrahal, Z. (2015). Open University Learning Analytics dataset. UCI Machine Learning Repository. DOI: 10.24432/C5KK69.

Le Quy, T., Roy, A., Iosifidis, V., Zhang, W., and Ntoutsi, E. (2022). A survey on datasets for fairness-aware machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(3):e1452. DOI: 10.1002/widm.1452.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35. DOI: 10.1145/3457607.

Patil, V., Ghalme, G., Nair, V., and Narahari, Y. (2021). Achieving fairness in the stochastic multi-armed bandit problem. *Journal of Machine Learning Research*, 22(174):1–31. Available at:`http://jmlr.org/papers/v22/20-704.html`.

Pessach, D. and Shmueli, E. (2022). A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 55(3):1–44. DOI: 10.1145/3494672.

Roh, Y., Lee, K., Whang, S. E., and Suh, C. (2021). Fairbatch: Batch selection for model fairness. DOI: 10.48550/arXiv.2012.01696.

Sáez, J. A. (2022). Noise models in classification: Unified nomenclature, extended taxonomy and pragmatic categorization. *Mathematics*, 10(20):3736. DOI: 10.3390/math10203736.

Silva, I. O. e., Soares, C., Sousa, I., and Ghani, R. (2024). Systematic analysis of the impact of label noise correction on ml fairness. In Liu, T., Webb, G., Yue, L., and Wang, D., editors, *AI 2023: Advances in Artificial Intelligence*, pages 173–184, Singapore. Springer Nature Singapore. DOI: 10.1007/978-981-99-8391-9_14.

Stefano, P. G. D., Hickey, J. M., and Vasileiou, V. (2020). Counterfactual fairness: removing direct effects through regularization. *CoRR*, abs/2002.10774. DOI: 10.48550/arXiv.2002.10774.

Wang, J., Liu, Y., and Levy, C. (2021). Fair classification with group-dependent label noise. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 526–536, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3442188.3445915.

Wu, S., Gong, M., Han, B., Liu, Y., and Liu, T. (2022). Fair classification with instance-dependent label noise. In Schölkopf, B., Uhler, C., and Zhang, K., editors, *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, pages 927–943. PMLR. Available at:`https://proceedings.mlr.press/v177/wu22b.html`.

Xia, X., Liu, T., Han, B., Wang, N., Gong, M., Liu, H., Niu, G., Tao, D., and Sugiyama, M. (2020). Part-dependent label noise: Towards instance-dependent label noise. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 7597–7610. Curran Associates, Inc. Available at:`https://proceedings.neurips.cc/paper_files/paper/2020/file/5607fe8879e4fd269e88387e8cb30b7e-Paper.pdf`.

Yeom, S. and Fredrikson, M. (2020). Individual fairness revisited: Transferring techniques from adversarial robustness. *CoRR*, abs/2002.07738. DOI: 10.24963/ijcai.2020/61.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. (2013). Learning fair representations. In *International conference on machine learning*, pages 325–333. PMLR. Available at:`https://proceedings.mlr.press/v28/zemel13.pdf`.