How Can Fairness Be Quantified For Applications In AI?

Grazia Obuzor
Department of Electrical and Electronic Engineering (EEME)
University of Bristol (Year 2)
Bristol, England
tk22151@bristol.ac.uk

***Abstract*—This project focuses on quantifying fairness in AI using metrics such as Statistical Parity, Equalized Odds, and Intersectional Statistical Parity; these implementations are accessible from a public GitHub repository*. These metrics offer a framework for evaluating the AI model's performance across various demographic groups. By implementing and analysing these fairness definitions within Python scripts, the study explores the trade-off between precision and recall, assesses the models discriminatory abilities through ROC curves, and utilises AUC as a fairness measure. Additionally, Chi-Square tests are employed to detect biases among groups and separate confusion matrices observed to calculate error rates, providing an evaluation of fairness in AI models as well as how their performance accuracy is affected. Intersectional fairness (via demographic parity) seems to produce ideal outcomes, suggesting there should be further research exploiting this definition.**
Code: https://github.com/xgpo-2226/Quantifying-Fairness-In-AI

***Index Terms*—implementation, MLP, fairness, TPR, FPR, parity, equalised odds, underrepresented demographics, accuracy score**

Nomenclature

AI	Artificial Intelligence
AUC	- Area Under the Curve
EQO	Equalised Odds (Fairness)
FNR	False Negative Rate
FPR	False Positive Rate
IXF	Intersectional FairnesS
ML	Machine Learning
MLP	Multi-Layer Perceptron
PR	Precision Recall
ROC	Receiver Operating Characteristic
SP	Statistical Parity (Fairness)
TNR	True Negative Rate
TPR	True Positive Rate

I. Introduction

Fairness in artificial intelligence (AI) is a complex and multifaceted concept that encompasses both ethical and technical factors. Ethically, fairness in AI is universally understood to involve handling personal data in ways that people reasonably expect and avoiding actions that lead to negative discrimination against individuals or groups [1]. Ensuring fairness in AI systems requires a nuanced understanding of biases, equitable treatment across different demographic groups, and maintaining transparency and accountability throughout the AI development pipeline. Technically, there is no universal approach to fairness in AI, as many mathematical definitions exist, and different types are used depending on the model and desired outcomes. Historically, controversial algorithms like COMPAS which was used for predicting re-offenders, raised concerns about fairness, due to their "black box" nature and adverse effects. It was found to be biased against certain demographic groups, particularly African Americans, leading to higher rates of false positiveswhere individuals are incorrectly predicted to re-offend. With no publicly available documentation offering insight as to how the algorithm worked. These issues call into question: Are existing definitions of fairness sufficient for most applications? Is there a single definition that stands out, or are new approaches needed to address the diverse and evolving applications of AI? The importance of fairness in AI cannot be overstated, and so the study of ethical AI practices is increasingly evolving - set out to ensure that systems do not perpetuate bias. Transparency in AI models allows practitioners and stakeholders to understand how data is being used, fostering trust and accountability when malfunctions or AI causes harm. However, balancing fairness with other key objectives, such as accuracy, is often challenging. Enhancing fairness may require trade-offs with performance metrics, depending on the specific application. The balance between accuracy and fairness can shift based on the desired outcome; in some cases, greater emphasis may be placed on fairness and equitable representation, while accuracy may take precedence in others. For instance, in healthcare, ensuring equal treatment outcomes across demographics is critical, whereas in hiring, avoiding biases based on gender, race, or other attributes is paramount. This report briefly considers the role fairness plays in various applications, ranging from healthcare predictions and natural language processing (NLP) to hiring practices and facial recognition systems, and how these considerations influence the choice of fairness definitions. Primarily, it focuses on experimenting with and analysing the Python implementation of some of the most widely recognised definitions: Equalized Odds, Statistical Parity, and Intersectional (Parity) Fairness; concluding with suggested ways to best quantify fairness.

II. Literature Review: Quantifying Fairness In AI

The following section introduces some key descriptions and mathematical expressions for existing fairness definitions in the field of AI; here, both the three definitions implemented for experimental analysis and two other definitions mentioned later in the report are discussed. Each definition addresses different concerns in achieving fairness.

A. STATISTICAL PARITY

Statistical parity (SP) (or demographic parity) ensures the predicted outcomes of a model are independent of any sensitive attributes, such as race or gender, meaning that the probability of a positive outcome should be the same across all groups defined by the sensitive attribute. For binary classification, with a sensitive attribute S and an outcome *P*, SP can be mathematically expressed as:

$$P(Y = 1|\ S = 0) = P(Y = 1|\ S = 1) \tag{1}$$

This formula indicates that the probability of a positive outcome (e.g., being hired) should be the same whether the individual belongs to the group S = 0 or S = 1 (e.g., male or female). Furthermore, where it is conditional, SP may include an additional condition probability:

$$P(Y = 1|S = 0, F = f) = P(Y = 1|S = 1, F = f) \tag{2}$$

where f is the probability of a positive instance of another feature (e.g., degree class in a hiring model).

SP is most suitable for applications such as in hiring practices, college admissions, or loan approvals, where practitioners (e.g. HR Recruitment) often see it as important to ensure that an AI model's outcomes are equally distributed for all different demographic groups rather than individuals. It is particularly useful in scenarios where representation across groups is more important than individual fairness. However, one of its limitations is that it may not account for nuanced backgrounds found in different groups. Some warn that SP can sometimes lead to reverse discrimination, where the decision process favours a group merely to equalise outcomes, potentially disregarding merit-based considerations.

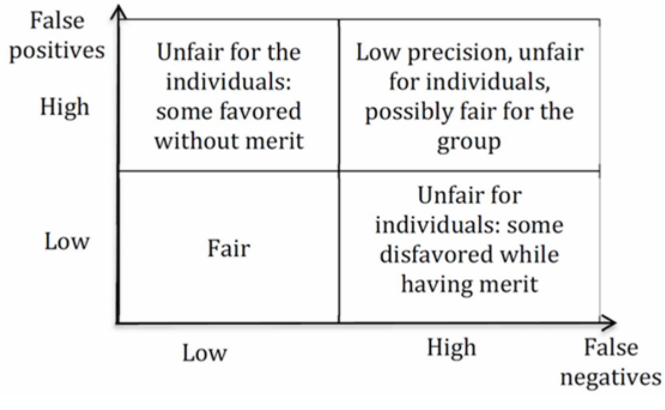

False positives
High
Low
Unfair for the individuals: some favored without merit
Low precision, unfair for individuals, possibly fair for the group
Fair
Unfair for individuals: some disfavored while having merit
Low
High
False negatives

Fig. 1. Fairness at the individual or group level

B. EQUALISED ODDS

Equalized odds (EQO) ensures a machine learning model's predictions are equally accurate across different groups defined by a sensitive attribute, regardless of the actual outcomes; the model's true positive rate (TPR) and false positive rate (FPR) are the same for all groups. It can be mathematically expressed as:

$$P(\hat{Y} = 1|S = 0, Y = y) = P(\hat{Y} = 1|S = 1, Y = y), \quad y \in 0, 1 \tag{3}$$

where Y is the predicted outcome, $\hat{Y}$ is the actual outcome, and S is the sensitive attribute. This definition ensures that the models accuracy is consistent across groups, both in terms of detecting true positives and avoiding false positives - the importance of which is explained in the documentary ”Coded Bias” when demonstrating the biases in profiling and facial recognition.

EQO is particularly applicable where it is important to ensure that a model does not disproportionately favour or harm any group. It is commonly applied in settings like criminal justice, healthcare, and facial recognition, where both false positives and false negatives can have severe consequences for individuals. Its primary challenge to implementation is when there are inherent differences between groups. Balancing the true positive and false positive rates across groups might require trade-offs that reduce the overall accuracy or utility of the model. However, the definition was selected for experimental analysis here as developers often use it to implement group fairness (in terms of both the benefits and harms of the models predictions).

C. INTERSECTIONAL FAIRNESS (THROUGH STATISTICAL PARITY)

Intersectional fairness (IXF) refers to the concept of ensuring fairness across multiple, intersecting demographic groups (e.g. race, gender, and socioeconomic status combined) rather than focusing on each sensitive group attribute separately. As a still emerging definition of fairness in ML studies, there are many existing experimental approaches to implementing IXF. The decision to pick statistical parity for analysis was again due to its simplicity in implementing the same positive rate (in the binary classification case) across intersecting groups. It can be mathematically expressed as:

$$P(\hat{Y} = 1|A = a) = P(\hat{Y} = 1|A = b) \tag{4}$$

where A is the protected attribute (e.g., gender, race), and $\hat{Y}$ is the predicted outcome. This equation ensures that the probability of a positive prediction is the same regardless of the intersecting demographic group observed *a* or *b*.

IXF is best used in scenarios where the goal is to ensure equal representation or treatment across different demographic groups, such as in hiring processes, loan approvals, or school admissions. Its main challenge: lack of data often results in some intersectional groups not being accounted for affecting how well the model's overall IXF value is measured. Due to the inherently niche intersecting groups that exist, it is a type of fairness that is the most difficult to realise among the three definitions chosen for experimentation. Moreover, where there may be sufficient representation of different demographic groups, GDPR rules restrict the types of sensitive data that can be collected from individuals. Despite this, it is valuable to explore IXF in experimental analysis because few studies document how IXF compares to the previous two definitions for measuring fairness in AI.

D. CALIBRATION

Calibration fairness refers to ensuring that all instances, for all subgroups in a sample, assigned a probability *p* of a positive outcome have a proportion *p* of actual positive outcomes. For example, if a model predicts a 70% chance of a positive outcome overall, then in 70% of such cases, the actual outcome should be positive, across all demographics. It is best used when using a predictive model, ensuring that the predicted probabilities are trustworthy and consistent across different groups - usually used in risk assessments, medical diagnoses, and any scenario where probability estimates are critical for decision-making. It can also flexibly be used in the in/post-processing stage of the AI development pipeline.

However, it is difficult to achieve a high EQO measurement while also using calibration, meaning one technique must be prioritised and it is advised to implement both definitions to achieve fairness in one model. A model can be calibrated but still exhibit disparities in FPR and FNR values, which could perpetuate unfairness - it does not address fairness across groups directly. And so is not explored in the experimental analysis since it would be incompatible with the second implementation used.

E. COUNTERFACTUAL FAIRNESS

Counterfactual fairness ensures that an AI program's predictions remain unchanged in a hypothetical world where a protected attribute (e.g., race, gender, socioeconomic status) is altered while keeping all other factors constant. It ensures that predictions should not depend on the sensitive attribute itself, but only on factors that are not causally related to the attribute. It can be mathematically described as:

$$\hat{Y}(A \leftarrow a) = \hat{Y}(A \leftarrow b) \tag{5}$$

where $\hat{Y}$ is the predicted outcome, and A is the protected attribute, meaning the model is considered fair if the prediction when A is altered from a to b is the same. This makes this approach to achieving fairness most applicable in where practitioners wish to ensure that decisions are not influenced by factors correlated with protected attributes, such as in hiring, credit scoring/loans, and law enforcement.

Yet this definition relies on having a correct and complete causal model of the relationships between variables, which can be difficult to determine in practice. Moreover, it may not capture all dimensions of fairness if the protected attribute indirectly influences outcomes through other unprotected attributes of a sample (such as occupation or hobbies implying gender), which makes it more complicated than what could be implemented in the experimental analysis within the duration of this project.

F. FAIRNESS THROUGH UNAWARENESS

Fairness through unawareness is the principle that a model is fair if it does not explicitly use protected attributes (e.g., race, gender) as features. It supports the idea the model cannot directly discriminate based on these factors this way. This approach is best, in theory, applied to preventing discrimination in contexts where sensitive information is prohibited from or unable to be collated.

As with some previous definitions, it ignores the possibility of proxy variables potentially leading to indirect discrimination. It also does not address biases present in the training data, making it less robust than the other definitions selected for experimental analysis - where it is difficult to track these proxy features especially.

III. METHODOLOGY

Through the implementation and evaluation of a Multi-Layer Perceptron (MLP) model with several variations for the three chosen fairness definitions, it is possible to begin to explore how fairness can be quantified in AI systems. Key decisions in the program revolve around the choice of model, functions, data handling, and several fairness and performance analysis methods, as explained in this section.

AI Model Selection: Multi-Layer Perceptron (MLP) — The MLP model was chosen because it is the simplest form of neural network that can be implemented in Python. It is a type of model with a fully connected network where each layer's nodes are connected to every node in the next layer. Allowing for easier optimisation. Moreover, Python was useful as the interpreter as it has many libraries, such as TensorFlow, PyTorch, and scikit-learn, provide numerous tools (such as schedulers and various loss functions) for optimising and analysing MLP models.

Dataset Selection: The dataset was downloaded from the UCI repository; UCI ID=697 (or the dataset entitled: "Predict Students' Dropout and Academic Success") sample [3] was ideal as it is well-documented, with no missing data, and uses categorical labels. This choice simplifies pre-processing and ensures that the focus remains on the model and fairness analysis rather than wasting time data-cleaning and manipulating. Categorical data allows for more well-defined groups to exist, which is necessary for assessing performance and fairness. One of the real-world implications of the models realised in this project is in identifying students who may need more support to improve retention or determining academic scholarship and funding allocation.

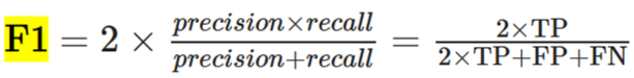
Train-Test Split and Sensitive Attributes: The dataset is divided into training and testing sets to evaluate the model's performance. Sensitive attributes, gender/nationality/educational special needs, are selected according to ethical guidelines [4]. The split is determined such that the model is trained on a subset of the data and tested on another and avoids under- or over-fitting.

Three Scripts for Fairness Metrics: The project employs three different scripts, each corresponding to a different fairness metric. These scripts are designed to integrate fairness evaluations into the loss function of the MLP model. Albeit, all scripts are programmed to measure the three fairness statistics post-testing stage, and are also all subject to the same performance analysis. The fairness metrics likely include statistical parity, equalized odds, and intersectional fairness (through statistical parity), experimentally quantify how well the model treats different demographic groups in predicting how likely an individual is to have attained a certain level of education.

ReLU Activation Function: The ReLU (Rectified Linear Unit) function is used in the hidden layers of the MLP; it handles non-linearity in a model while being computationally efficient. Loss Function and Adversarial Approach: The loss function incorporates an adversarial approach where the model is penalised for high loss, particularly when there are disparities between groups, in terms of the fairness statistics calculated in each respective script.

Evaluation Metrics: The project uses several metrics to evaluate the model, including

F1 Score: Balances precision and recall, providing a measure that accounts for both false positives and false negatives.

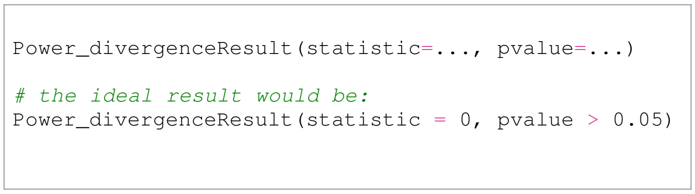


$$\mathbf{F1} = 2 \times \frac{precision \times recall}{precision + recall} = \frac{2\times\mathrm{TP}}{2\times\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}$$

Fig. 2. F1-Score Equation

as in (Hicks et al., 2022) [5]. For a good model, a value of 0.7+ is ideal.

Chi-Squared Distance: Used for categorical data to measure the distance between observed and expected frequencies (for accuracy evaluation), and between distributions of all demographic groups of protected attributes (which helps in fairness evaluation); ideally, the distance should be negligible, and since the result is given in the form



```
Power_divergenceResult(statistic=..., pvalue=...)

# the ideal result would be:
Power_divergenceResult(statistic = 0, pvalue > 0.05)
```

The three possible labels are converted to ”1-hot” values to create a 1D vector thats easier to handle. The chi-squared statistic indicates the absolute distance between the distributions of predictions between different demographic groups, while a p-value less than 0.05 indicates statistical significance in their difference. This is used instead of the Wasserstein distance which is less suitable for discrete classification since it does not assume that there are several bins within the distribution of labels in a dataset.

Confusion Matrix: Provides insight into the model's performance across different classes, revealing disparities in how different groups are treated, i.e. a visual representation of the respective implementation's TNR/TPRs, and FNR/FPRs as in (Split, 2024) [6].

ROC and PR Curves: Receiver Operating Characteristic (ROC) (Richardson et al., 2024) [7] curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various prediction threshold positions (boundary values used by the model to determine how to classify a label, e.g. perhaps 0.5 for a binary classifier). A perfect ROC curve has a right angle at the top left corner, indicating high sensitivity (TPR) and specificity (TNR); the Area Under the ROC Curve (AUC-ROC) is ideally 1.0, signifying perfect classification. Whereas the Precision-Recall (PR) curve displays the trade-off between precision (i.e. accuracy, rate of correctly predicted labels) and recall (i.e., sensitivity or TPR) for different threshold values. Ideally, the curve shows high precision and recall across various thresholds, with a square curve with a right angle to the right corner of the graph. Where there is perfect performance with no trade-off between precision or recall, the area under the graph is also 1.0.

Accuracy/Fairness trade-off —
Achieving a perfectly fair model may come at the cost of overall accuracy, and the decision to prioritize one over the other depends on the specific application and ethical considerations. Ultimately, accuracy is a pivotal metric to record for each implementation script [insert equation here] is measured and compared to see which one offers the best trade-off.

Running the Model (model hyperparameters): The model, on the computer used for testing, is run using 0 workers to ensure it runs as fast as possible. It is run multiple times (i.e. four times) to manually calculate average values and assess the consistency of the results. The remaining hyperparameters are set to align with standard values used in similar testing [8] for optimal results across the three scripts. Furthermore, schedulers and optimisers are used for further optimisation.

IV. Analysis and Discussion

In the Table of Results, the outputs of testing the three different implementations chosen are displayed. The possible class labels are referred to as: ”Dropout/Class 0,” ”Enrolled/Class 1” and ”Graduate/Class 2”. In this section, these findings are evaluated and compared to reach a conclusion to the overall project question.

A. Precision-Recall Curve

In the IXF implementation, the curves for ”Dropout,” ”Enrolled,” and ”Graduate” show a relatively similar pattern, with precision decreasing as recall increases. The ”Graduate” class has a higher precision at lower recall levels compared to the other classes, suggesting this model is more confident at predicting the graduate class and struggles to distinguish between the other two in comparison. Moreover, since the Enrolled class PR-Curve is vastly different to the other two curves, there is reason to believe that this model is biased towards Enrolled students, having less precision classifying them than other types of individuals.

B. ROC Curve Analysis

The ROC curve for each class shows the trade-off between True Positive Rate (TPR) and False Positive Rate (FPR). The black dotted line shows the least ideal ROC curve shape it indicates an essentially random predictor. The IXF model seems to have the best shaped curves, with their curves closest to a right angle towards the top left corner of the graph, and the highest true positive rates.

The IXF model also seems to have ROC-AUC values that are the closest to 1.0, ranging from 0.78 to 0.84, confirming that it has a superior predictive ability.

Being closely related to the concept of Equalized Odds, which ensures that the model's performance (in terms of TPR and FPR) is similar across different demographic groups, the ROC curve for the IXF model has a smaller range of ROC-AUC values, meaning less disparity across the class distributions followed by the SP model, and the EQO model performing the worst with the largest disparity in prediction error rates.

C. Fairness Statistics Results

From the numerical results, the IXF models overall accuracy on the test set is 75.18%, with an F1-score of 0.75, indicating balanced precision and recall (as indicated in its PR-Curve); consequently, this model has the lowest Hamming Loss of 0.248 (i.e. under 25% of incorrect classifications). The Intersectional Statistical Parity score is 0.09759, which demonstrates some disparity between groups, although the closest to zero than in analysis of the other two models. This is also true of the low Statistical Parity score (0.00232) for this model, indicating that the model does not significantly favour one demographic group over another in its predictions, an essential aspect of fairness. It is lower than in the intersectional case, revealing that there is more fairness between male/female groups generally than in more specific demographic groups, e.g. male/female individuals of specific nationalities with certain educational learning needs. The Equalized Odds Rate is 0.00000, indicating no (or likely negligible due to representation limits of values in Python) observed disparities in TPR and FPR between groups. Surprisingly, this result is even better than that of the EQO value derived in the EQO model itself.

D. Chi-Square Analysis

The Chi-Square tests for accuracy (goodness-of-fit tests) show poor performance across the groups, with large chi-squared statistics (i.e. distance between actual and predicted distributions) and <0.05 p-value results, affirming that the predicted values consistently fall outside of the range of the actual labels. However, the Chi-Square tests for fairness (homogeneity tests) across all groups yield p-values of 1.0, and much lower disparity in across all group distributions. This suggests that achieving high fairness compromises accuracy for all implementations.

E. Confusion Matrix

The error and precision rates are calculated for each class per model, with FPR and TPR values noted in the results table; the confusion matrix corresponding to the IXF model specifically is as follows:

Predicted/Actual	Dropout	Enrolled	Graduate
Dropout	283	96	32
Enrolled	43	325	55
Graduate	6	87	358

True Positives (TP):
Dropout (283): 283 instances of Dropout were correctly classified as Dropout.
Enrolled (325): 325 instances of Enrolled were correctly classified as Enrolled.
Graduate (358): 358 instances of Graduate were correctly classified as Graduate.
The high true positive counts across all classes suggest that the model generally performs well in accurately identifying the correct class for most instances.

False Positives (FP):
Dropout (96 + 32 = 128): The model incorrectly predicted 96 Enrolled and 32 Graduate instances as Dropout.
Enrolled (43 + 55 = 98): 43 Dropout and 55 Graduate instances were incorrectly predicted as Enrolled.
Graduate (6 + 87 = 93): 6 Dropout and 87 Enrolled instances were incorrectly predicted as Graduate.
The false positive rates vary, with Dropout having the highest number of false positives, suggesting that the model is more likely to misclassify instances as Dropout compared to Enrolled or Graduate.

False Negatives (FN):
Dropout (43 + 6 = 49): 43 Enrolled and 6 Graduate instances were mistakenly classified as Dropout.
Enrolled (96 + 87 = 183): 96 Dropout and 87 Graduate instances were mistakenly classified as Enrolled.
Graduate (32 + 55 = 87): 32 Dropout and 55 Enrolled instances were mistakenly classified as Graduate.
The false negative rates are notably higher for the Enrolled class, which could indicate a bias or a particular challenge in correctly identifying students who are enrolled.

True Negatives:
Dropout (1285 283 128 49 = 825).
Enrolled (1285 325 98 183 = 679).
Graduate (1285 358 93 87 = 747).
The true negative rates are notably higher than false positive

values, which further indicates good accuracy.

F. Accuracy vs. Fairness

In this experiment, the goal is to find the model with the accuracy and fairness measurements, rather than aiming for any specific balance one way or another.

Evidently, from the results that emerge, the IXF model seems to fulfil this aim the most; based on this, it seems that this fairness metric is the best performing amongst the three investigated. In practice, however, it may not always be ideal to prioritize representative predictions, especially when the primary objective is to maintain accuracy unaffected by fairness interventions, or when sensitive data required for tracking performance across specific groups is unavailable.

While the IXF model implemented in this project used statistical parity, several other approaches could offer different perspectives on IXF in AI. IXF through Equality of Opportunity focuses on ensuring that individuals across different demographic intersections have equal chances of achieving favourable outcomes, such as being admitted to a program or receiving a loan, i.e. this model would be adapted to ensure that the TPRs for ”Dropout,” ”Enrolled,” and ”Graduate” are consistent across all demographic groups, most likely via an algorithm to adjust the decision thresholds for each group. Alternatively, IXF through Individual Fairness, says that similar individuals should receive similar treatment by the AI system; so, the model should ensure fairness across multiple, overlapping identities rather than group-wise. A k-nearest neighbours (k-NN) [link] approach could be used to enforce that individuals with similar characteristics across different intersections are treated similarly. For example, a Black woman and a White woman with similar qualifications would have similar chances of being identified as enrolled or graduated. This method might create a more nuanced model that can mitigate more bias and avoid the need for more intersectional group data to train the model, but it could increase computational complexity and require more detailed and granular data.

Developers instead of practitioners often determine the trade-off between accuracy and fairness; the purpose of the AI model should be the main factor in this decision. For instance, in the educational context of the dataset used in these experiments, where the goal might be to identify students at risk of dropping out, fairness in predictions could be crucial to ensure that more support is offered to marginalised (i.e. niche intersectional) groups. Similarly, in healthcare, attaining better accuracy for specific subgroups, such as non-white disabled women, and ensure equitable outcomes, could require more of a focus on fairness; otherwise, for example, a diagnosis model could have a high accuracy score for men overall, but further analysis could reveal that for the sub-group of non-white men, predictions are consistently inaccurate. In this case, the model could artificially appear fairly accurate, but by being principally incorrect for more intersectional demographics, rendering the model ineffective and causing disproportionally more harm for particular stakeholders.

When considering other fairness metrics like calibration, counterfactual fairness, and fairness through unawareness, it’s essential to select the metric that best aligns with the model’s intended purpose. Calibration ensures that predictions are accurate across different groups, which is crucial in contexts like criminal justice, where overrepresentation of certain groups due to unresolved biases could have severe consequences. In this context, however, it may not be unfair for a particular group to be overrepresented in a specific label if the individuals are all correctly identified. Fairness Through Unawareness, which involves omitting sensitive attributes from the model, simplifies the model design and prevents direct discrimination. In the current implementation, it could reduce the complexity of managing multiple sensitive attributes, and even avoid the challenge of lack of representation of particular groups in a sample though, proxy variables could still have a negative effect if not resolved. Implementing counterfactual fairness is computationally intensive and requires a well-defined causal model. In cases where the causal relationships are complex or unknown, this method may result in inaccuracy, and counterintuitively introduce more bias towards groups whose contextual attributes may have an unresolved influence on the outcome.

The question remains whether these definitions are sufficient to cover all scenarios. Relying on a single metric, such as IXF, might not be adequate. Although IXF fairness generates superior quantitative values amongst others, with the method even including better equalised odds outcomes than implementing an EQO method itself, the choice of fairness metric in AI development should also consider the model’s goals. Whether the objective is to ensure equal distribution of predictions (opportunity), or make error rates consistent across groups (misclassifications, rejections, etc.) the appropriate metric must align with the models purpose to achieve truly fair outcomes.

V. LIMITATIONS AND FUTURE CONSIDERATIONS

One of the primary challenges is posed by the dataset’s size and diversity. The disparities between the samples across demographic groups, introduces significant biases that can impact model training and the fairness of outcomes. Future research could employ resampling techniques to mitigate these imbalances. For instance, the Bias Mimicking Sampling method [8], which can either oversample certain features or groups, or synthetically generate this data.

Future research should also aim to compare the outcomes of these different fairness metrics using datasets from various domains, such as healthcare, hiring, and justice, to determine their effectiveness in other contexts.

Another area for future exploration is the ethical involvement of stakeholders in the development and deployment of machine learning models. In the healthcare domain, for instance, it is crucial to consult with diverse stakeholders to ensure that the models developed do not perpetuate existing inequalities. It is essential that stakeholders are not exploited but rather receive fair outcomes, as highlighted in the Netflix documentary Coded Bias, where the inhabitants of an apartment block were forced to trial new facial recognition software to enter their home without proper consent or consideration of their rights (00:22:15).

Standardised frameworks that can be used to evaluate and implement fairness across different domains could be formulated in future research; for example, these frameworks would incorporate various fairness metrics and provide guidelines for their application in different contexts. The development of such frameworks would greatly assist developers and practitioners in ensuring that their models adhere to ethical standards and do not perpetuate biases.

VI. CONCLUSION

The objective of this study was to explore how fairness can be quantified in AI models and to assess whether existing and emerging fairness definitions sufficiently cover all aspects of fairness for diverse applications. By evaluating these definitions, the project aimed to uncover their effectiveness and limitations, particularly in the context of academic achievement.

Fairness metrics can ensure that AI models do not disproportionately disadvantage certain groups, thereby promoting equitable opportunity to attain desirable outcomes and be rejected or misclassified at a similar rate across different groups. The goal in this project was to quantify fairness by minimising the difference in model outcomes across different strata, aiming for distribution distances close to zero.

Looking forward, there is a need for more research on alternative and emerging fairness metrics, especially those that account for intersectionality and subgroup disparities, since this seemed the most promising definition from the experimental data, with regards to both high accuracy scores and equitable predictions amongst the intersectional groups.

One implication of this project is that it highlights the importance of collaborative data sharing in enhancing the representation of underrepresented groups. By pooling resources across institutions data scarcity can be addressed, ensuring that even sensitive information, which cannot always be directly collected, is considered. This approach is pivotal for intersectional analysis and encourages future research into methods for resolving variables that are often challenging to capture, for more comprehensive and fair AI models.

To conclude, fairness in AI is an evolving field that requires continuous research and adaptation. While existing definitions of fairness provide a solid foundation, they are not exhaustive enough to address the complex nature of fairness across all applications. The slight disparities observed in intersectional fairness metrics underscore the necessity for research into more nuanced and adaptive fairness measures. Ultimately, at present, balancing accuracy and fairness is crucial to ensuring AI models promote equitable outcomes for a specific use, determined by practitioners who would consider the specific context and objectives of their models.

REFERENCES

[1] Information Commissioner's Office. (n.d.). What about fairness, bias, and discrimination? ICO. Available at: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources (Accessed: 01 August 2024).

[2] RES.EC-001 exploring fairness in machine learning, fairness criteria: Exploring fairness in machine learning for international development: Edgerton center MIT OpenCourseWare. Available at: https://ocw.mit.edu/courses/res-ec-001 (Accessed: 01 August 2024).

[3] Realinho,Valentim, Vieira Martins,Mnica, Machado,Jorge, and Baptista,Lus. (2021). Predict Students' Dropout and Academic Success. UCI Machine Learning Repository. Available at: https://doi.org/10.24432/C5MC89. (Accessed: 12 July 2024).

[4] What personal data is considered sensitive? European Commission. Available at: https://commission.europa.eu/law/law-topic/data-protection/ (Accessed: 01 August 2024).

[5] Hicks, S.A. et al. (2022) On evaluation metrics for medical applications of Artificial Intelligence, Scientific reports. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8993826/ (Accessed: 03 August 2024).

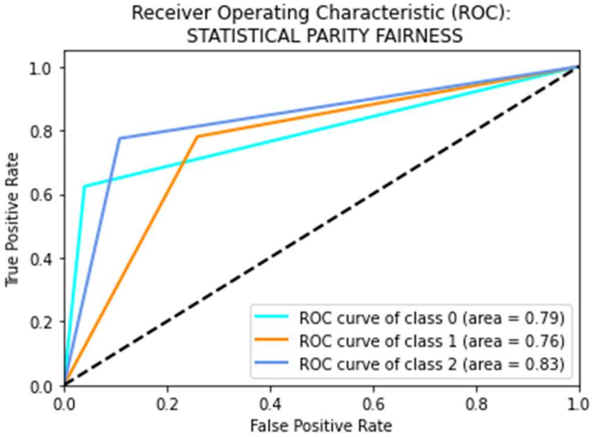
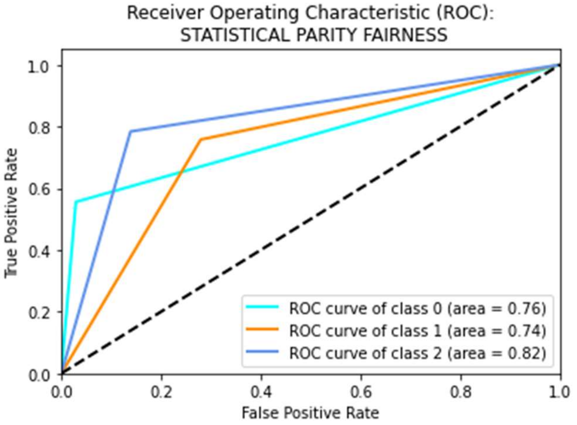
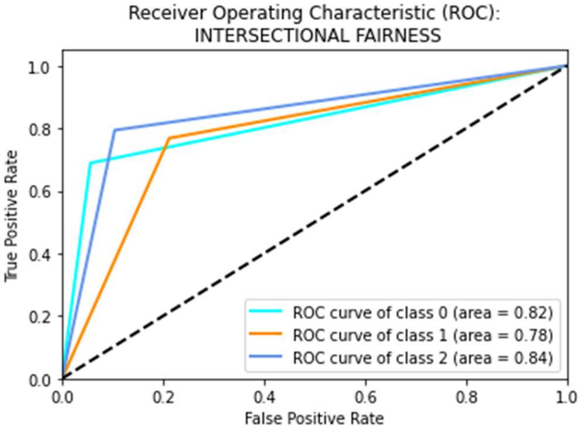
[6] False positive rate (2024) Split. Available at: https://www.split.io/glossary/false-positive-rate/#:∼:text=The%20true%20positive%20rate%20(TPR,as%20TN%2FTN%2BFP.https://www.split.io/glossary/false-positive-rate/ (Accessed: 03 August 2024).

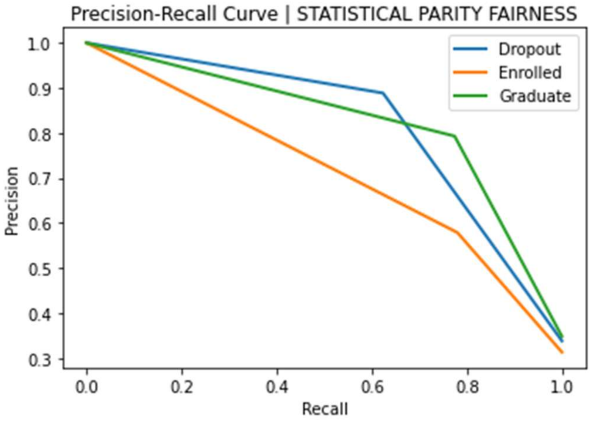
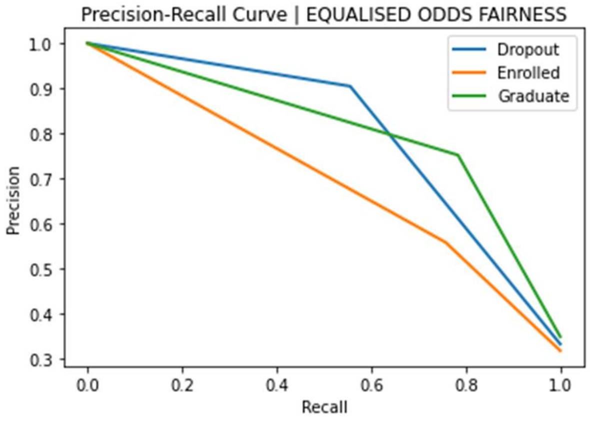
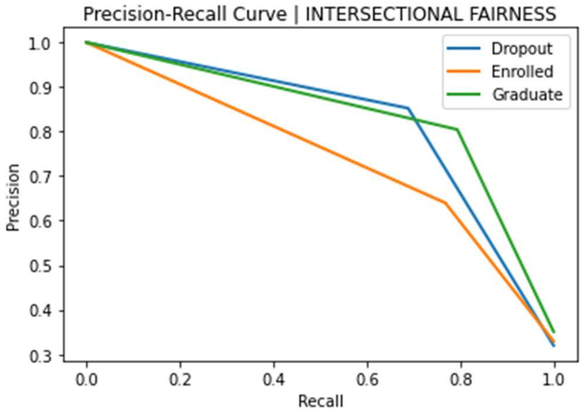
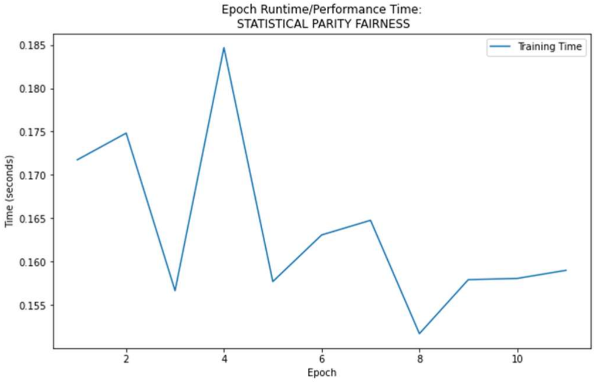
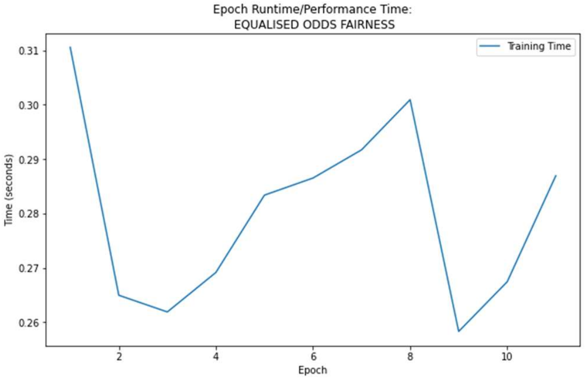
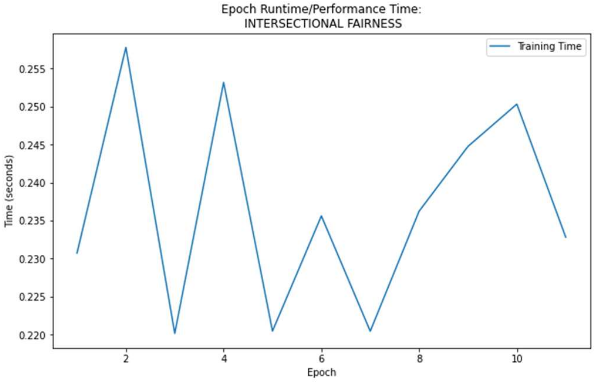
[7] Richardson, E. et al. (2024) The receiver operating characteristic curve accurately assesses imbalanced datasets, Patterns, 5(6), p. 100994. doi:10.1016/j.patter.2024.100994.

[8] https://doi.org/10.48550/arXiv.2209.15605

EXPERIMENTAL FINDINGS		
SP	**EQO**	**IXF**
Training time: 1.64 seconds Testing time: 0.07 seconds	Training time: 3.08 seconds Testing time: 0.10 seconds	Training time: 2.60 seconds Testing time: 0.05 seconds
Accuracy of the network on the test set: 72.518% Accuracy Score is: 0.7252	Accuracy of the network on the test set: 69.9494% Accuracy Score is: 0.6995	Accuracy of the network on the test set: 75.18% Accuracy Score is: 0.7518
Statistical Parity: 0.00116 Equalised Odds Rate: 0.00000 Intersection Statistical Parity: 0.13929	Statistical Parity: 0.00433 Equalised Odds Rate: 0.01282 Intersection Statistical Parity: 0.19368	Statistical Parity: 0.00232 Equalised Odds Rate: 0.00000 Intersectional Statistical Parity: 0.09759
F1-Score: 0.727	F1-Score: 0.69955	F1-Score: 0.75
Classification Report: precision recall f1-score support 0 0.89 0.62 0.73 869 1 0.58 0.78 0.66 805 2 0.79 0.77 0.78 895 accuracy 0.73 2569 macro avg 0.75 0.73 0.73 2569 weighted avg 0.76 0.73 0.73 2569	Classification Report: precision recall f1-score support 0 0.90 0.56 0.69 855 1 0.56 0.76 0.64 817 2 0.75 0.78 0.77 897 accuracy 0.70 2569 macro avg 0.74 0.70 0.70 2569 weighted avg 0.74 0.70 0.70 2569	Classification Report: precision recall f1-score support 0 0.85 0.69 0.76 411 1 0.64 0.77 0.70 423 2 0.80 0.79 0.80 451 accuracy 0.75 1285 macro avg 0.77 0.75 0.75 1285 weighted avg 0.77 0.75 0.75 1285
Hamming Loss: 0.2748151031529778	Hamming Loss: 0.3005060334760607	Hamming Loss: 0.2482490272373541
Confusion Matrix: [[542 266 61] [57 628 120] [11 191 693]]	Confusion Matrix: [[475 299 81] [47 619 151] [3 191 703]]	Confusion Matrix: [[283 96 32] [43 325 55] [6 87 358]]

The False Positive Rate for the label 'Dropout' is: 0.1669218989280245 The True Positive Rate for the label 'Dropout' is: 0.8885245901639345 The False Positive Rate for the label 'Enrolled' is: 0.1192722371967655 The True Positive Rate for the label 'Enrolled' is: 0.5788018433179724 The False Positive Rate for the label 'Graduate' is: 0.1191740412979351 The True Positive Rate for the label 'Graduate' is: 0.7929061784897025	The False Positive Rate for the label 'Dropout' is: 0.18590998043052837 The True Positive Rate for the label 'Dropout' is: 0.9047619047619048 The False Positive Rate for the label 'Enrolled' is: 0.13561643835616438 The True Positive Rate for the label 'Enrolled' is: 0.5581605049594229 The False Positive Rate for the label 'Graduate' is: 0.11872705018359853 The True Positive Rate for the label 'Graduate' is: 0.7518716577540107	The False Positive Rate for the label 'Dropout' is: 0.13431269674711438 The True Positive Rate for the label 'Dropout' is: 0.8524096385542169 The False Positive Rate for the label 'Enrolled' is: 0.12612612612612611 The True Positive Rate for the label 'Enrolled' is: 0.639763779527559 The False Positive Rate for the label 'Graduate' is: 0.11071428571428571 The True Positive Rate for the label 'Graduate' is: 0.8044943820224719
ROC AUC Score (Macro-Averaged): 0.79516 ROC AUC Score (Micro-Averaged): 0.79389	ROC AUC Score (Macro-Averaged): 0.77489 ROC AUC Score (Micro-Averaged): 0.77462	ROC AUC Score (Macro-Averaged): 0.81300 ROC AUC Score (Micro-Averaged): 0.81381
Measuring ACCURACY with CHI-Square Chi-Square Test for Male Group: Power_divergenceResult(statistic=69.46164125729344, pvalue=8.252703628929117e-16) Chi-Square Test for Female Group: Power_divergenceResult(statistic=113.675042936631, pvalue=2.0690837088240816e-25)	**Measuring ACCURACY with CHI-Square** Chi-Square Test for Male Group: Power_divergenceResult(statistic=148.79960156611546, pvalue=4.8817672014003115e-33) Chi-Square Test for Female Group: Power_divergenceResult(statistic=143.41642175653803, pvalue=7.203104383990399e-32)	**Measuring ACCURACY with CHI-Square** Chi-Square Test for Group 1: Power_divergenceResult(statistic=84.48719724594494, pvalue=4.50648803480805e-19) Chi-Square Test for Group 2: Power_divergenceResult(statistic=52.87518422562475, pvalue=3.298371966046547e-12) Chi-Square Test for Group 3: Power_divergenceResult(statistic=49.82441334446297, pvalue=1.5162335782324605e-11)

		Chi-Square Test for Group 4: Power_divergenceResult(statistic=50.69258558554168, pvalue=9.82301717518648e-12) ... Chi-Square Test for Group 18: Power_divergenceResult(statistic=47.689185843606616, pvalue=4.4098606976756404e-11) Chi-Square Test for Group 19: Power_divergenceResult(statistic=47.811158058876536, pvalue=4.1489570669738326e-11)
Measuring FAIRNESS with CHI-Square Chi-Sqaure Statistic Across All Groups: 2.17521 P-Value Across All Groups: 0.33702	**Measuring FAIRNESS with CHI-Square** Chi-Sqaure Statistic Across All Groups: 5.27568 P-Value Across All Groups: 0.07152	**Measuring FAIRNESS with CHI-Square** Chi-Sqaure Statistic Across All Groups: 5.06755 P-Value Across All Groups: 1.00000
		

Precision-Recall Curve | STATISTICAL PARITY FAIRNESS Dropout Enrolled Graduate Precision Recall 1.0 0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.0 0.2 0.4 0.6 0.8 1.0	Precision-Recall Curve | EQUALISED ODDS FAIRNESS Dropout Enrolled Graduate Precision Recall 1.0 0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.0 0.2 0.4 0.6 0.8 1.0	Precision-Recall Curve | INTERSECTIONAL FAIRNESS Dropout Enrolled Graduate Precision Recall 1.0 0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.0 0.2 0.4 0.6 0.8 1.0
Epoch Runtime/Performance Time: STATISTICAL PARITY FAIRNESS Training Time Time (seconds) Epoch 0.185 0.180 0.175 0.170 0.165 0.160 0.155 2 4 6 8 10	Epoch Runtime/Performance Time: EQUALISED ODDS FAIRNESS Training Time Time (seconds) Epoch 0.31 0.30 0.29 0.28 0.27 0.26 2 4 6 8 10	Epoch Runtime/Performance Time: INTERSECTIONAL FAIRNESS Training Time Time (seconds) Epoch 0.255 0.250 0.245 0.240 0.235 0.230 0.225 0.220 2 4 6 8 10