Discovering Trade-offs in Fairness and Accuracy: A Multi-Objective Approach

Arsh Chowdhry[a,1,2], **Aishwaryaprajna** [b,3], **Apurva Narayan**[c,4] **and Peter R Lewis** [d,5]

[a]Ontario Tech University, International Centre for Applied System Science for Sustainable Development, CA
[b]University of Exeter, UK
[c]University of Western Ontario, CA
[d]Ontario Tech University, CA

Abstract. In recent years, there has been a shift from focusing exclusively on the accuracy of machine learning systems to a more holistic and human-centered approach that includes privacy, fairness, transparency and more. Many of these dimensions are often considered to conflict with each other. For example, there can be a trade-off between the accuracy and fairness of a predictive model. In fairness analysis, the aim is to establish that machine learning models do not discriminate based on protected or sensitive characteristics such as race, gender, age, or religion. In practice there are many alternative notions of fairness, some of which themselves may not be mutually compatible. In this paper, we explore this relationship between accuracy and different notions of fairness using German Credit dataset, where training a model using standard techniques has been shown to lead to biased predictions. We explore the trade-off between accuracy and six different fairness metrics using a multi-objective training approach, which aims to maximize both accuracy and fairness. Our results show that in certain cases, there exists a trade-off between accuracy and different notions of fairness. In these cases, the multi-objective approach provides a set of models that balance the trade-off in different ways. Further, in other cases, the approach does not lead to a trade-off, instead giving rise to a model that is both accurate and fair simultaneously, when this was not achieved using a single-objective approach. Therefore, we show that by explicitly targeting fairness during training, decision makers can have access to a range of models that might meet their accuracy and fairness requirements. Moreover, we also show that a multi-objective approach identifies situations where an assumed trade-off between fairness and accuracy need not exist.

1 Introduction

The positive impact of AI is undeniably profound, revolutionizing various aspects of our lives, including healthcare, business, and daily interactions [31]. AI's ability to deliver innovative solutions efficiently and accurately has significantly contributed to enhancing decision-making and problem-solving capabilities [11]. Nevertheless, from a socio-technical perspective, conventional machine learning systems encounter numerous challenges, with ethical considerations in AI being a prominent concern [24]. The growing reliance on AI has led researchers to closely examine issues related to the safety, privacy, and fairness of AI systems [41].

Despite extensive research on fairness in AI, the concept remains complex and multifaceted, continually raising the fundamental question: what defines fairness? Fairness can be defined differently across disciplines. Mulligan et al. [35] explore the concept of fairness across various disciplines, including law, social science, quantitative fields, and psychology. The paper [35] helps us to understand that within different disciplines, fairness definition can differ. While individuals may have different perspectives on what constitutes fairness, establishing a unified definition for fairness in both the real and algorithmic worlds remains a complex task [43, 3]. Human prejudice is a natural characteristic [19], and intelligent algorithms aim to mitigate this behavior by enabling machines to make human-like decisions without such biases. Despite significant advancements in AI that have greatly simplified our lives in various domains, including healthcare, education, finance, and agriculture [25], there are ongoing concerns regarding privacy, fairness, and transparency in AI that require further attention [26]. Ferrara defines fairness as the absence of bias or discrimination in AI systems [18]. Due to its subjective and contextual nature, the concept of fairness poses the challenge of selecting an appropriate definition, necessitating both quantitative and qualitative understanding. Quantitative understanding relates to the statistical measurements conducted on data, while qualitative understanding is contextual and dynamically changes according to specific needs and problems.

Claims regarding AI bias are not merely speculative; substantial evidence indicates the presence of bias in AI systems due to factors such as race, color, ethnicity, or religion [32]. For instance, in 2015, Amazon acknowledged that its recruitment algorithm exhibited gender-related bias. The algorithm penalized applicants whose resumes included words associated with women [14]. Since the algorithm was trained on resumes collected over a decade, it reflected a greater percentage of male applicants, leading to the rejection of female candidates [14]. In response, companies have begun implementing gender decoder tools to analyze and suggest changes to address gender disparities [13]. The application of AI extends beyond recruitment algorithms. AI is extensively used in

[1] Email: arsh.chowdhry@ontariotechu.net
[2] https://www.icasssd.org/
[3] Email: aishwaryaprajna@exeter.ac.uk
[4] Email: apurva.narayan@uwo.ca
[5] Email: peter.lewis@ontariotechu.ca

US courtrooms to predict the future behavior of criminals. The Correctional Offender Management Profiling for Alternative Sanctions (COMPASS) system assesses the likelihood of criminals reoffending [42]. In 2016, reports indicated that these algorithms exhibited bias against black defendants, incorrectly flagging black Americans as high risk and white Americans as low risk, resulting in black Americans being labeled as recidivists at nearly twice the rate of white Americans [1]. Efforts to address this included incorporating a "race-neutral" component to ensure similar performance across racial groups [4]. Evidence of AI bias is also present in the US healthcare system, where algorithms have been found to be racially biased against black patients, requiring them to be more gravely ill than white patients to receive equivalent care [47]. Researchers have addressed such issues by including subgroup analysis to identify and mitigate bias in AI models [48].

These real-world examples underscore the importance of considering fairness in AI systems. Ensuring fairness not only enhances the accuracy and sustainability of AI systems but also contributes to achieving broader societal goals [5] While conducting fairness analysis, a protected or sensitive attribute refers to a demographic category for which non-discrimination needs to be established, ensuring that the outputs of models do not discriminate against individuals based on characteristics such as race, gender, age, or religion. Let us suppose there are two applicants: Nathan (male applicant) and Zahra (female applicant). They both are requesting a loan amount of $40000 for 12 months. They both exhibit a positive checking account and have a credit account opened at the bank. The credit scores show that they have duly paid their credits to the bank. They are both skilled employees with more than two years of experience. Even though they possess the same criteria on paper and should have received the loan, the AI model did not approve Zahra's loan while approving Nathan's request. This raises questions regarding the model being susceptible to bias.

Integrating ethical attributes into AI systems presents several challenges due to the complexity of combining qualitative, subjective, and contextual factors. Addressing both accuracy and fairness simultaneously is particularly challenging. To tackle this issue, we frame accuracy and fairness as a Multi-Objective Optimization problem. This approach is adopted because accuracy and fairness represent conflicting objectives; improving one typically impacts the other. Our primary contribution is to explore the trade-offs between accuracy and various fairness metrics through a Multi-Objective Optimization approach. This framework aims to balance both accuracy and fairness, demonstrating that this approach can improve outcomes in both dimensions. Our findings reveal that, in some cases, a trade-off between accuracy and fairness exists, while in others, it is possible to achieve models that are both accurate and fair simultaneously. We validated our approach using several real-world datasets. The results indicate that the Multi-Objective framework can identify solutions that balance both accuracy and fairness in some instances, while in others, a trade-off between these objectives is necessary.

2 Related Work

AI-driven decision-making systems assist humans in various decision-making tasks such as hiring, lending loans, healthcare, and many more. These classification systems operate by learning patterns and rules from human-collected data and then making predictions on new, unseen data. Any classification task will adhere to these steps to enhance the model's overall performance iteratively. However, this process is not without challenges. The model must also be carefully evaluated to ensure that it does not introduce or perpetuate biases present in the historical data, which could affect fairness in loan approval decisions. These biases can emerge at every stage of the process, from data collection to analysis, training, and prediction [36]. Implementing such ethical practices is essential to mitigate potential discrimination or harm, particularly in high-stakes applications. To understand different types of fairness, a wide literature is available, differentiating into three broad categories: Group fairness, Individual fairness, and Counterfactual fairness [49]. Table 1 briefly describes the three broad categories.

Fairness Type	Definition	Examples
Group Fairness	Ensures AI system does not disproportionately benefit or harm particular group	Demographic Parity, Equalized Odds, Equal Opportunity, Disparate Impact
Individual Fairness	Similar treatment is done with similar individuals, regardless of the group.	Causal Discrimination, Fairness Through Unawareness, Fairness Through Awareness
Counterfactual Fairness	Ensures that AI decisions are not influenced by certain sensitive attributes that could cause discrimination	Causal Reasoning: Counterfactual fairness, No unresolved discrimination, No proxy discrimination, Fair Inference

Table 1: Fairness in AI can be broadly classified into three categories: Group, Individual, and Counterfactual fairness.

An important note is that achieving one type of fairness does not guarantee another. Given the complexity and multifaceted nature of fairness, the criteria for fairness can vary based on specific needs and contexts. In the context of analyzing fairness in ML systems, Verma and Rubin [45] discuss 20 different notions of fairness measurements in a classification model. These 20 definitions can be broadly sub-categorized into three categories: Statistical measures, Similarity-based measures, and Causal Reasoning. The paper illustrates the different notions of fairness in a human-interpretable way using a well-established case study based on a public credit scoring dataset. Research in fairness in AI is a growing field, which brings this study relatively new and rapidly evolving. In the article, Mitchellc et al.[34] points out the choices and assumptions made by researchers leading to fairness issues in the decision making procedure. Furthermore, highlighting choices and assumptions which may lead to fairness concerns. As ML systems have extensive application across various realms, including healthcare, education, transportation and many more, Pessach and Shmueli [37] provides an overview of how to identify, analyse, and improve the algorithmic fairness in the classification tasks. The fairness enhancing mechanisms includes; prepossessing, intraining, and post-processing mechanisms through which a model can achieve reasonable accuracy and fairness. The paper surveys to understand the pros and cons of ML systems, providing an idea regarding the use of different notions of fairness definitions, as they are domain-specific and delve into the emerging research in algorithm fairness. Pre-processing techniques tries to remove the source of bias before model training by excluding the features correlated with the protected attributes, reducing the attributes of input features in the dataset. However, this results in data loss, which simultaneously removes and introduces bias [10, 40]. Employing techniques such as sampling, reweighing, massaging [27] or even selecting some attributes in the dataset [39] as a data preprocessing step in ML systems can influence bias output. An ex-

ample of fairness notion here would be fairness through unawareness [29] where sensitive attributes are discounted from the data before training, influencing bias in the data. Post-processing techniques modify the predictions after training the algorithm. Following this mechanism, Pleiss et al. [38] investigate the relationship between calibration and error rates. In the paper, calibration refers to the number of instances when the prediction aligns with the model's output, and error rates refer to the model's ability to perform similarly across different subgroups in the dataset. Their study illustrates that calibration is only compatible with the fairness notion of equal opportunity (false negative rate). The authors emphasize that when calibration complies with the false negative rate, any algorithm used is no more effective, and the predictions are generated by chance, concluding that calibration and error rate constraints are incompatible objectives. Pre-processing and post-processing techniques include a straightforward method of changing the data before and after the training process. In the in-training technique, modifications are made while training the algorithm itself [46]. Generative adversarial networks (GANs) are widely used in the training process, leveraging Neural Network architecture for classifying and giving fair results during adversarial attacks. Zhang et al. [51] propose adversarial networks for mitigating bias in the algorithm. Apart from GANs, the idea of using multi-objective optimization algorithms has recently been proposed. FOMO [30] uses the NSGA-II algorithm to optimize fairness and accuracy in the model. Unlike other optimization techniques, such as gradient descent, multi-objective algorithms do not have any regularization techniques. The method proposes a novel meta-model that maps protected attributes to sample weights, helping to optimize the weights.

Hardt et al. [22], proposes a combination of two metrics known as Equalized Odds, which uses post-processing technique to enhance fairness and accuracy in the models. A toolkit developed by Microsoft[2], tries to use the reduction approach in the training process for optimizing fairness in systems. The paper includes demographic parity and equalized odds. Foulds et al.[20] suggests the inclusion of a definition which satisfies three criteria of intersectionality, privacy, and economic guarantees. The approach utilizes the fairness cost as regularization to balance the trade-offs between accuracy and fairness.

Despite substantial research aimed at mitigating bias in ML systems, the application of a multi-objective perspective remains relatively underexplored. In our research, we investigate the relationship between fairness metrics and model accuracy. Through this approach, we seek to offer valuable insights into the relationship between fairness and accuracy in ML models.

3 Measuring Fairness

In this section, we will be discussing about the statistical measures for analysing fairness. Statistical notions are fundamental for understanding fairness in classification-based machine learning models [45]. These notions are defined using a confusion matrix and probability metrics. A confusion matrix is a tool for evaluating the performance of a classification algorithm, where the rows represent the predicted classes and the columns represent the actual classes. Each class in the matrix consists of positive and negative values. In our research, we have chosen six fairness notions from the thirteen statistical definitions available. These selections are based on their use of confusion metrics to quantify fairness within machine learning models. While other statistical fairness definitions might be intuitive, they often lack formalization, which is crucial for integration with clarity metrics. Table 2 summarizes the mathematical interpretation of these definitions expressed in terms of statistical measurements and probabilities. The notations used in the equations can be interpreted as:

- $\hat{y}$ = Predictive decision for an individual, 1 indicates a positive outcome, and 0 refers to a negative outcome.
- G = Protected or the sensitive attribute in the dataset, 1 indicating protected attribute, and 0 indicating unprotected attribute.
- Y = Actual classification decision for an individual.

Since treatment equality measures the ratio of errors in the model, there is no probability equation in the table. To understand the intuitive meaning of these definitions, we explain six fairness notions with their real-world application.

No.	Fairness Definition	Probability Equation	Statistical Meaning
1.	Group Fairness [16]	$P(\hat{y}=1 \mid G=0) = P(\hat{y}=1 \mid G=1)$	Positive Prediction
2.	Predictive Parity [12]	$P(Y=1 \mid \hat{y}=1, G=0) = P(Y=1 \mid \hat{y}=1, G=1)$	False Negative Rate (FDR)/ Positive Predictive Value
3	Predictive Equality [12]	$P(\hat{y}=1 \mid Y=0, G=0) = P(\hat{y}=1 \mid Y=0, G=1)$	False Positive Rate (FPR)/ True Negative Rate (TNR)
4	Equal Opportunity [12]	$P(\hat{y}=0 \mid Y=1, G=0) = P(\hat{y}=0 \mid Y=1, G=1)$	False Negative Rate (FNR)/ True Positive Rate (TPR)
5	Overall Accuracy [6]	$P(\hat{y}=Y, G=0) = P(\hat{y}=Y, G=1)$	Accuracy
6	Treatment Equality [6]	-	Ratio of False Negative and False Positive

Table 2: Six fairness notions are considered in the paper, with their Mathematical interpretation through probabilities and their statistical significance.

3.1 Group Fairness (Statistical Parity) [16]

Ensuring group fairness involves assessing the positive predictions in the dataset to ensure equal treatment across all demographic groups. The intuition behind this definition is that, after training the algorithm and generating predictions, positive predictions should be consistent across different subgroups in the dataset.

3.2 Predictive Parity (Outcome Test) [12]

Unlike group fairness, this definition not only considers the algorithm's predictions but also compares them with the actual outcomes of the data. The intuition behind this definition is that the algorithm's predictions should exhibit equal accuracy with the actual outcomes among different demographic groups.

3.3 Predictive Equality (False positive error rate balance) [12]

This fairness metric calculates those instances while the model's prediction is positive, but the actual outcome is negative. Intuitively, predictive equality arises when the algorithm erroneously

predicts the positive instances for a particular demographic group that belongs to a negative class. A society may face severe repercussions for having a high false positive rate.

3.4 Equal Opportunity (False negative error rate balance) [12]

Inverting the notations of Predictive Parity by swapping predictions of the model ($\hat{y}$) from 1 to 0 and actual outcomes (Y) from 0 to 1 promotes the calculation of Equal Opportunity. Intuitively, Equal Opportunity represents those instances when the algorithm mistakenly predicts a negative outcome despite the actual outcome being positive.

3.5 Overall Accuracy Equality [6]

Overall accuracy measures the model's capability to correctly classify those instances when the predictions align with the actual outcome. It evaluates the model's probability of correctly predicting positive outcomes when the actual outcome is positive and its probability of correctly predicting negative outcomes when the actual outcome is negative. Overall accuracy can be crucial while conducting fairness analysis, particularly within the healthcare sector.

3.6 Treatment Equality [6]

Treatment equality calculates the ratio of false negatives and false positives for the different demographic groups in the dataset. The uniqueness of calculating the ratio of errors makes this definition fall apart from all the other definitions discussed above. It measures the magnitude of errors of the model across different demographic groups.

4 Dataset

For our baseline experiment, we employed the German Credit Dataset [23], consisting of 1000 loan applicant records and 20 attributes. The goal of conducting the fairness analysis on the German dataset is to check whether males and females are granted loans equally or whether there is discrimination based on the sex of an individual. Personal status and sex attributes in the original dataset are classified into five classes: 1. Divorced/separated males, 2. Divorced/separated/married females, 3. Single males, 4. Married/widowed males, and Single females. Out of 1000 records in the dataset, there was no instance reported for single female applicants, leading to bias in the dataset. To overcome this issue, we created two columns from the personal status and sex attributes and applied one hot encoding that originated two new attributes: i. Sex, where 1 indicates females, and 0 indicates males, ii. The personal status column assigning 1 to married applicants and 0 to single applicants. Finally, the dataset comprised 57 variables and 1000 entries, containing an additional attribute describing the outcome, whether the applicant has a good or a bad credit score.

5 Baseline Experimental Study

Fairness analysis aims to establish that ML systems do not discriminate based on sensitive attributes in a dataset, such as sex, race, or religion. For conducting fairness analysis, our methodology utilizes the standard logistic regression model for the classification problem [44]. The rationale behind choosing logistic regression for training the model is that we reimplemented the fairness analysis on a well-established case study [45] and chose the same algorithm for training the model. The next step involves splitting training data as 80% and test data as 20%. We applied a 10-fold cross-validation technique [7] to measure the model's accuracy. Finally, we obtain the predictions and calculate the model's performance, resulting in a final accuracy of 75%.

In our experiments, we are interested in conducting a fairness analysis on sex-related descriptions and whether male and female applicants are treated equally regarding the loan approval model. To calculate fairness, we need the following four attributes: i) Predicted outcomes of the model, ii) Actual outcomes in the dataset, iii) sensitive attribute (females), and iv) insensitive attribute (males). Once we have these four attributes, we can calculate the fairness metrics for the sensitive attribute. In general, during statistical fairness analysis, the difference between the probabilities associated with the sensitive and insensitive attributes is compared against a threshold to determine the presence of significant bias. Such a threshold is often an arbitrary choice. Verma and Rubin [45] employ an arbitrary threshold of 0.06 for the statistical fairness measurements. Indicating that the definition would be deemed unfair if there is a discrepancy of 6% or more. Determining a meaningful threshold for a model is domain-specific. For example, in the healthcare sector, we aim to have a threshold as small as possible so that the difference between the sensitive attributes is minimal, whereas, in other scenarios, the threshold may vary. For statistical validity, we introduced statistical tests in our analysis by conducting a "Student's t-test" [33] on the sensitive attributes of the six fairness notions. By implementing these tests, we understand whether an observed disparity represents a significant difference between the two groups or whether the result we got is of random chance, avoiding the need to define an arbitrary threshold. By conducting the baseline experiments, we understand if there is any disparity between the sensitive group and the insensitive group in a given problem.

Definition	Male	Female	Fairness Difference	Statistical T-test
Group Fairness	0.808±0.038	0.691±0.050	**0.118±0.063**	**3.60E-14**
Predictive Parity	0.803±0.036	0.765±0.048	0.056±0.048	**0.0012**
Predictive Equality	0.433±0.096	0.550±0.083	**0.178±0.097**	**4.11E-06**
Equal Opportunity	0.897±0.035	0.828±0.056	**0.079±0.050**	**8.86E-07**
Overall Accuracy	0.765±0.030	0.726±0.035	.049±0.035	**2.26E-05**
Treatment Equality	0.504±0.246	0.774±0.518	**0.454±0.401**	**0.014**

Table 3: Results for our baseline experiment. Columns two and three represent the fairness analysis conducted on sex attributes for each definition. The fourth column represents the standard way of analysing fairness in ML models, where if the difference between the male and female applicants is more than a threshold, the definition is deemed unfair. The bold instances in the column represent when a significant difference between male and female groups persists. Column five illustrates the p-values from the t-test results.

Our results are presented in Table 3, where we calculate the six different notions of fairness for each male and female applicant

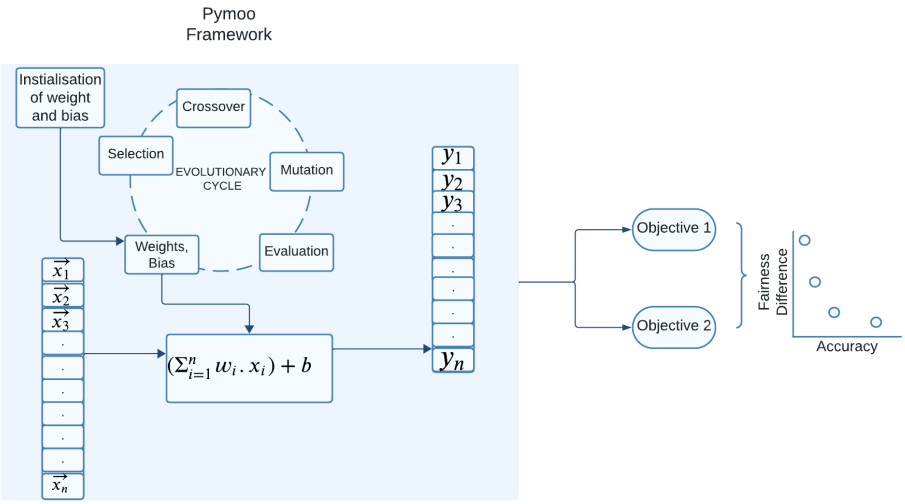

Figure 1: Overview of Multi-objective framework.

in the dataset. To ensure that the model's performance is consistent and reliable, we ran the model thirty times, assessing fairness for each subgroup and presenting the mean and standard deviation for the fairness metric. This gives us an average of fairness and indicates how much fairness varies. In Table 5.2, columns second and third provide information regarding each subgroup's mean and standard deviation. In Table 3, columns second and third provide information regarding each subgroup's mean and standard deviation. Choosing an arbitrary threshold of 6% would mean that any variation of more than six percent between the male and female columns would be considered unfair. Column four illustrates the difference between the two subgroups for each fairness notion. For comparison with Verma and Rubin [45], the following two fairness notions, Predictive Equality and Overall Accuracy, would satisfy the fairness criteria as they do not exceed the 6% threshold, while the remaining fairness notions are considered unfair.
In the fairness analysis, we used t-tests to establish a general and uniform criterion, demonstrating the presence of bias in all cases. This experiment demonstrates that bias is present in the training of ML models and reveals that as the accuracy of the model increases, the disparity between the male and female groups increases significantly, indicating the presence of bias in the ML systems. In the next section, we will implement our multi-objective approach on the same dataset and analyse the tradeoffs between the two conflicting objectives.

6 Multi-Objective Approach

Recent studies have shown that optimizing accuracy in ML systems may prevent us from including human-centric approaches in the decision-making process [45]. Thus, the aim is to build accurate machines replicating human-centric behaviours such as fairness, transparency, explainability, etc. Multi-objective optimization tries to simultaneously optimize multiple conflicting objective functions within a given problem [28]. Since these objectives often conflict with one another, achieving all of them at once can be challenging. This approach is utilized in various domains, such as healthcare and education [21]. Typically, focusing on one objective results in a trade-off with another. The goal of multi-objective optimization is to identify a set of solutions that form the Pareto-optimal front, representing the best trade-offs among the conflicting objectives. Problems that involve optimizing multiple objectives are referred to as multi-objective optimization problems.The mathematical definition for a multi-objective optimization problem can be expressed as follows:

$$\begin{aligned} &\text{minimize/maximize} && f_n(x),\ n = 1,\ldots,N. \\ &\text{subject to} && g_i(x) \le 0,\ i = 1,\ldots,m. \\ & && h_j(x) = 0,\ j = 1,\ldots,p. \end{aligned}$$

Where:

- n ≥ 2, number of objectives
- $f_n(x)$, objective functions
- $g_i(x)$, inequality constraints
- $h_i(x)$, equality constraints

In the equation, $x = [x_1, x_2, x_3, \ldots. x_N]$ represents the vector of decision variables, where each point in the decision space specifies a set of values for these variables. Conversely, the objective space is where the objectives (criteria to be optimized) are evaluated, with each point in this space corresponding to the objective values achieved by a solution in the decision space. To analyze trade-offs between solutions, we map solutions from the decision space to the objective space. $f_n(x)$ represents the objective function that needs to be minimized or maximized. In a multi-objective optimization framework, solutions must satisfy two types of constraints: inequality and equality constraints. Inequality constraints restrict the acceptable values within the decision space, representing regions where the constraint function is less than or equal to zero. Equality constraints, on the other hand, define functions that must be exactly equal to zero. The multi-objective optimization problem aims to identify the optimal set of solutions that effectively balance conflicting objectives while satisfying the constraints.
For this paper, we implemented our experiments on a Python framework that utilizes an NSGA-II algorithm [15] known as pymoo [8]. The Pymoo framework can support many Evolutionary Algorithms [17], such as NSGA-II, MOEAD, PSO, etc. The framework requires three arguments: 1. problem definition, 2. algorithm used (NSGA-II algorithm), and 3. the termination criteria. For configuring the multi-objective approach in pymoo, we used the following parameters: i) Population size: 40, Offspring size: 10, Crossover Probability: 0.9, Crossover: Simulated binary crossover

[50], Mutation: Polynomial Mutation [17], Number of Iterations: 50000 and eliminate duplicates: True. Evolutionary Algorithms [17] are inspired by natural selection, which utilizes the evolutionary process of crossover, selection, and mutation to generate new solutions in each generation. The figure 1 provides an overview of our multi-objective experiment aimed at uncovering the trade-offs between fairness and accuracy in the training process. In any classification task within a machine learning model, certain parameters are crucial for guiding the algorithm's learning process during training. These parameters include weights (w) and bias (b). Weights indicate the contribution of each input feature to the output, while bias aids in prediction even when input features are zero. The number of weights corresponds to the number of input features in the dataset. In a simple logistic regression model, we represent the linear combination of weights and bias with the input data as $z = (\Sigma_{i=1}^{n} w_i.x_i)+b$ [52]. This linear combination is then fed into the sigmoid function to make predictions. Subsequently, the algorithm employs optimization techniques, such as gradient descent, to minimize the difference between the prediction and the actual output. In our research, we altered this process so that the weights and bias evolve directly using the Evolutionary Algorithm. The next step is to define our objectives for accuracy and fairness notions. Finally, we utilize the pymoo framework and the NSGA-II algorithm to convert our problem statement into a multi-objective problem. This approach helps us to examine the trade-offs between accuracy and fairness for the sensitive attributes in the dataset.

7 Results

In this section,we explore the trade-offs between accuracy and each of the six different fairness notions using a multi-objective approach to maximise accuracy and fairness.
In this experiment, we seek to understand the following two cases:

- In the first scenario, if we do not identify a trade-off, it suggests that both objectives are optimized simultaneously, indicating that we have achieved maximum fairness and accuracy in a model.
- Conversely, if we discover a trade-off between the two objectives, it provides decision-makers with a range of models that can meet their requirements and aids in identifying which fairness notion they should prioritize in the given context.

Each Figure in our experiment plots a set of non-dominated solutions for accuracy and fairness differences between sensitive and insensitive groups. A model can be interpreted as fair in the figures when the fairness difference between the sensitive groups is close to 0. Since pymoo can be adopted only for minimisation problems, we minimise our problem while defining our problem statement, which leads to negative accuracy in the figures. For example, -0.78 indicates 78% of model accuracy. As discussed in the above section, sex is considered a sensitive attribute, and the goal is to minimise the fairness difference while achieving maximum accuracy.
Achieving both objectives in a problem is challenging since accuracy and fairness conflict with each other [9]. While running the experiments, our results discovered some cases where a trade-off between accuracy and fairness metrics exists. In other cases, the approach does not lead to a trade-off; instead, it gives rise to a model that is both accurate and fair simultaneously, which was not achievable using a single-objective approach.

To understand the significance of each figure, let us first examine Figure 2, where we analyse the trade-off between group fairness and accuracy. As the accuracy improves, the fairness difference increases from 0% to a maximum of 2%. The increase in fairness difference is insignificant with an increase in accuracy, suggesting the fairness difference to be insignificant. Thus, we identify no substantial trade-off between the two objectives. It implies that the model is accurate and fair simultaneously. Figure 3 presents a similar trend for equal opportunity. As accuracy increases, the fairness difference increases from 0% to a maximum of 0.07%. This shows that the trade-off is not significant, satisfying our problem statement. When we consider predictive equality, we find that as accuracy improves from 70% to 72%, there is a significant increase in the difference in predictive equality from 0% to a maximum of 7%. It suggests the presence of a trade-off between the two objectives. Figure 4 illustrates that up to a certain point, there is a minimal increase in fairness difference with accuracy. However, beyond that point, as accuracy increases, the fairness difference also increases. This complex relationship between accuracy and fairness in the context of predictive equality highlights the need for decision-makers to explicitly target fairness during training to meet their accuracy and fairness requirements. For Figure 5, as accuracy increases, the increase in fairness difference is negligible. Once again, this shows no significant trade-off between the two objectives and satisfying our problem statement. Similarly, for Figure 6, the increase in accuracy is inversely proportional to the increase in fairness difference. As accuracy increases from 75% to 77%, fairness difference increases from 0% to 1%. The evidence shows no trade-off between accuracy and predictive parity. In Figure 7 as well, as accuracy increases, fairness difference rises from 0% to a maximum of 2%. It indicates that as the accuracy increases, the rise in fairness difference is negligible, indicating the absence of a significant trade-off between accuracy and treatment equality. Therefore, achieving high accuracy and fairness in the model. A noteworthy observation across all the graphs is that despite maintaining a uniform population size for all the observations, the number of instances on the non-dominated set varies. For example, 5 displays only three points in the objective space, whereas Figure 3 displays a higher density of instances. Due to the complexity of the problem, a varied number of points are on the Pareto front in each graph.

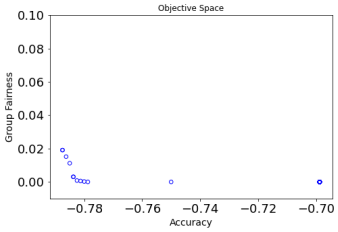

Figure 2: No significant trade-off is discovered between Group fairness and Accuracy

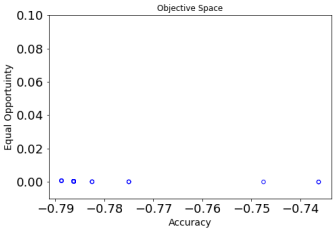

Figure 3: No significant trade-off is discovered for Equal Opportunity and Accuracy

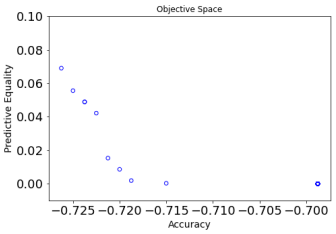

Figure 4: A significant trade-off is discovered for Predictive Equality and Accuracy

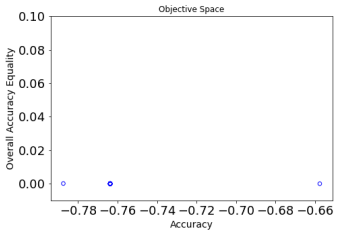

Figure 5: No significant trade-off is discovered for Overall Accuracy and Accuracy

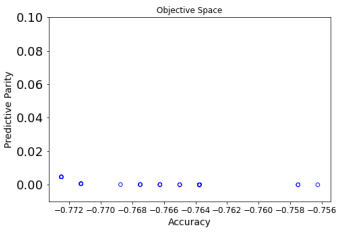

Figure 6: No significant trade-off is discovered for Predictive Parity and Accuracy

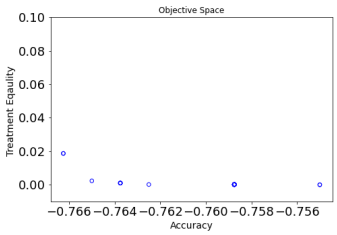

Figure 7: No significant trade-off is discovered for Treatment Equality and Accuracy

Fairness Definition	Baseline Experiment	Multi- objective Experiment
Group Fairness	✗	✓
Predictive Parity	✗	✓
Predictive Equality	✗	Trade-off
Equal Opportunity	✗	✓
Overall Accuracy	✗	✓
Treatment Equality	✗	✓

Table 4: Summary of our multi-objective results, comparing them with the baseline results from single-objective training across all fairness notions. ✗indicating the presence of bias, while ✓indicates when the model is fair.

8 Analysing The Results

Table 4 compares our multi-objective results with our baseline results across six fairness notions. The table records our decisions for six fairness notions for both experiments, giving a holistic review regarding our baseline experiments and multi-objective approach. When we implemented fairness analysis on the standard ML models and conducted the statistical tests on the experiments, the results showed that these techniques led to biased predictions. By incorporating a multi-objective approach in the training process, apart from predictive equality, which suggests a tradeoff between accuracy and fairness, all the other five definitions satisfy our objectives, suggesting a fair and accurate model. It shows that by employing fairness in a multi-objective approach, when a trade-off was assumed, we discovered no tradeoff between the two conflicting objectives, satisfying our problem, which was missing in the single-objective approach. The tradeoffs for predictive equality against accuracy provide information regarding the relationship between the two objectives. This indicates that even by explicitly targeting fairness during training to meet their accuracy and fairness requirements, we could not optimize the two objectives for predictive equality. Therefore, this allows decision-makers to understand the relationship better and prioritize the need.
Predictive equality calculates those instances when the model's prediction is positive, but the actual outcome is negative. Our experiment studies the relationship between accuracy and fairness; the tradeoff provides vital information in the context of predictive equality. A tradeoff for predictive equality means that improving the model's accuracy (i.e., its ability to correctly predict whether an individual will be approved for the loan) comes at the cost of increasing disparity in the false positive rates between the sensitive group (e.g., women) and the insensitive group (e.g., men). This indicates a bias in the model where increasing accuracy may lead to one group being unfairly misclassified more often.

9 Discussion

We began by conducting a fairness analysis by training a standard machine learning algorithm on the loan prediction dataset, which revealed the presence of biased predictions. Subsequently, we applied our framework to enhance fairness and accuracy with respect to sex attributes in the German Credit datasets. These baseline experiments underscore the existence of bias in conventional ML models. While several approaches are available to mitigate bias, our research aims to demonstrate that Multi-Objective Optimization can also effectively balance accuracy and fairness. Instead of comparing the efficiency of different methods, our focus is on providing an alternative solution that can aid decision-makers in aligning model outcomes with their specific requirements.

10 Conclusion

Our research investigates the relationship between six distinct fairness notions and accuracy within a multi-objective framework. The fairness analysis performed on the dataset reveals existing disparities in machine learning systems. Given that accuracy and fairness are often conflicting objectives, our goal is to maximize both within this framework. In certain cases, achieving both objectives simultaneously proves challenging, leading to a trade-off between fairness and accuracy. Experiments conducted on the German Credit Dataset demonstrate that our approach successfully optimizes both objectives in most scenarios. These findings provide decision-makers with a deeper understanding of the relationship between accuracy and fairness across various problem domains, allowing them to select models that align with their specific accuracy and fairness criteria. There are several points we would like to investigate in our future work. First, we intend to examine the relationship between fairness and accuracy with respect to other sensitive attributes, such as race, ethnicity, and religion, beyond gender. Second, future work will focus on analyzing fairness across different demographic groups rather than on individual attributes. Third, We aim to investigate the relationships among various definitions of fairness. Understanding these relationships will help us select conflicting fairness metrics and explore trade-offs between different fairness definitions within a multi-objective framework. As research on fairness in AI is still emerging, addressing these points will contribute significantly to advancing our understanding of fairness in machine learning systems and will be valuable to the AI research community.

References

[1] Can the criminal justice system's artificial intelligence ever be truly fair?, May 2021. URL https://massivesci.com/articles/machine-learning-compas-racism-policing-fairness/.

[2] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach. A reductions approach to fair classification. In *International conference on machine learning*, pages 60–69. PMLR, 2018.

[3] E. Amigó, Y. Deldjoo, S. Mizzaro, and A. Bellogín. A unifying and general account of fairness measurement in recommender systems. *Information Processing & Management*, 60(1):103115, 2023.

[4] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias. In *Ethics of data and analytics*, pages 254–264. Auerbach Publications, 2022.

[5] A. M. Astobiza, M. Toboso, M. Aparicio, and D. López. Ai ethics for sustainable development goals. *IEEE Technology and Society Magazine*, 40(2):66–71, 2021.

[6] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44, 2021.

[7] D. Berrar et al. Cross-validation., 2019.

[8] J. Blank and K. Deb. Pymoo: Multi-objective optimization in python. *Ieee access*, 8:89497–89509, 2020.

[9] M. Buyl and T. De Bie. Inherent limitations of ai fairness. *Communications of the ACM*, 67(2):48–55, 2024.

[10] F. Calmon, D. Wei, B. Vinzamuri, K. Natesan Ramamurthy, and K. R. Varshney. Optimized pre-processing for discrimination prevention. *Advances in neural information processing systems*, 30, 2017.

[11] S. Chakraborty. Ai and society: A case study on positive social change, 2024. https://www.techuk.org/resource/ai-and-society-a-case-study-on-positive-social-change.html: :text=The

[12] A. Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[13] K. Crawford and R. Calo. There is a blind spot in ai research. *Nature*, 538(7625):311–313, 2016.

[14] J. Dastin. Amazon ditches ai recruiting tool that didn't like women, 2018. URL https://globalnews.ca/news/4532172/amazon-jobs-ai-bias/.

[15] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE transactions on evolutionary computation*, 6(2):182–197, 2002.

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

[17] A. E. Eiben, J. E. Smith, A. Eiben, and J. Smith. What is an evolutionary algorithm? *Introduction to evolutionary computing*, pages 25–48, 2015.

[18] E. Ferrara. Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci*, 6(1):3, 2023.

[19] H. D. Fishbein. *Peer prejudice and discrimination: The origins of prejudice*. Psychology Press, 2014.

[20] J. R. Foulds, R. Islam, K. N. Keya, and S. Pan. An intersectional definition of fairness. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pages 1918–1921. IEEE, 2020.

[21] N. Gunantara. A review of multi-objective optimization: Methods and its applications. *Cogent Engineering*, 5(1):1502242, 2018.

[22] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

[23] H. Hofmann. Statlog (German Credit Data). UCI Machine Learning Repository, 1994. DOI: https://doi.org/10.24432/C5NC77.

[24] M. M. Islam. Ethical considerations in ai: Navigating the complexities of bias and accountability. *Journal of Artificial Intelligence General science (JAIGS) ISSN: 3006-4023*, 3(1):2–30, 2024.

[25] M. R. Islam, M. U. Ahmed, S. Barua, and S. Begum. A systematic review of explainable artificial intelligence in terms of different application domains and tasks. *Applied Sciences*, 12(3):1353, 2022.

[26] K. Kalimeri and I. Tjostheim. Artificial intelligence and concerns about the future: A case study in norway. In *International Conference on Human-Computer Interaction*, pages 273–284. Springer, 2020.

[27] F. Kamiran and T. Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33(1):1–33, 2012.

[28] A. Konak, D. W. Coit, and A. E. Smith. Multi-objective optimization using genetic algorithms: A tutorial. *Reliability engineering & system safety*, 91(9):992–1007, 2006.

[29] M. J. Kusner, J. Loftus, C. Russell, and R. Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.

[30] W. G. La Cava. Optimizing fairness tradeoffs in machine learning with multiobjective meta-models. *arXiv preprint arXiv:2304.12190*, 2023.

[31] S. Makridakis. The forthcoming artificial intelligence (ai) revolution: Its impact on society and firms. *Futures*, 90:46–60, 2017.

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

[33] P. Mishra, U. Singh, C. M. Pandey, P. Mishra, and G. Pandey. Application of student's t-test, analysis of variance, and covariance. *Annals of cardiac anaesthesia*, 22(4):407, 2019.

[34] S. Mitchell, E. Potash, S. Barocas, A. D'Amour, and K. Lum. Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 8:141–163, 2021.

[35] D. K. Mulligan, J. A. Kroll, N. Kohli, and R. Y. Wong. This thing called fairness: Disciplinary confusion realizing a value in technology. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–36, 2019.

[36] E. Ntoutsi, P. Fafalios, U. Gadiraju, V. Iosifidis, W. Nejdl, M.-E. Vidal, S. Ruggieri, F. Turini, S. Papadopoulos, E. Krasanakis, et al. Bias in data-driven artificial intelligence systems—an introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3):e1356, 2020.

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

[38] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger. On fairness and calibration. *Advances in neural information processing systems*, 30, 2017.

[39] A. U. Rehman, A. Nadeem, and M. Z. Malik. Fair feature subset selection using multiobjective genetic algorithm. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, pages 360–363, 2022.

[40] C. Saplicki. Amazon ditches ai recruiting tool that didn't like women, 2023. URL https://medium.com/ibm-data-ai/fairness-in-machine-learning-pre-processing-algorithms-a670c031fba8.

[41] B. Shneiderman. Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered ai systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 10(4):1–31, 2020.

[42] J. Skeem and J. Eno Louden. Assessment of evidence on the quality of the correctional offender management profiling for alternative sanctions (compas). *Unpublished report prepared for the California Department of Corrections and Rehabilitation. Available at: https://webfiles. uci. edu/skeem/Downloads. html*, 2007.

[43] J. Sylvester and E. Raff. What about applied fairness? *arXiv preprint arXiv:1806.05250*, 2018.

[44] M. Tranmer and M. Elliot. Binary logistic regression. *Cathie Marsh for census and survey research, paper*, 20, 2008.

[45] S. Verma and J. Rubin. Fairness definitions explained. In *Proceedings of the international workshop on software fairness*, pages 1–7, 2018.

[46] M. Wan, D. Zha, N. Liu, and N. Zou. In-processing modeling techniques for machine learning fairness: A survey. *ACM Transactions on Knowledge Discovery from Data*, 17(3):1–27, 2023.

[47] D. R. Williams and R. Wyatt. Racial bias in health care and health: challenges and opportunities. *Jama*, 314(6):555–556, 2015.

[48] S. Yan, H.-t. Kao, and E. Ferrara. Fair class balancing: Enhancing model fairness without observing sensitive attributes. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 1715–1724, 2020.

[49] M. B. Zafar, I. Valera, M. Gomez 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*, pages 1171–1180, 2017.

[50] F. A. Zainuddin, M. F. Abd Samad, and D. Tunggal. A review of crossover methods and problem representation of genetic algorithm in recent engineering applications. *International Journal of Advanced Science and Technology*, 29(6s):759–769, 2020.

[51] B. H. Zhang, B. Lemoine, and M. Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.

[52] X. Zou, Y. Hu, Z. Tian, and K. Shen. Logistic regression model optimization and case analysis. In *2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT)*, pages 135–139, 2019. doi: 10.1109/ICCSNT47585.2019.8962457.