Multi-Objective Approach to Balance Fairness and Accuracy

Zahid Irfan*, Róisín Loughran*, Muhammad Adil Raja*, Fergal McCaffery*
*Regulated Software Research Center (RSRC),
Dundalk Institute of Technology (DkIT),
Dundalk, Ireland
Email: {zahid.irfan, roisin.loughran, adil.raja, fergal.mccaffery}@dkit.ie

***Abstract*—As Artificial Intelligence systems are being deployed in multiple domains, ensuring they exhibit fair and just behaviour is a critical challenge. Multi-objective optimization offers a robust framework for addressing this challenge by simultaneously optimizing conflicting objectives, such as fairness and accuracy. In this work, we leverage causal graphs to model dependencies and identify potential sources of bias. We evolve directed acyclic graphs that represent causal structures, optimizing them for fairness and accuracy using evolutionary computational methods. Our approach employs multi-objective optimization to explore trade-offs between these objectives, enabling the discovery of solutions that balance ethical considerations with performance. Experimental results demonstrate that the multi-objective framework effectively improves fairness while maintaining competitive accuracy alongside building causal graphs. This approach provides a scalable and interpretable solution for mitigating bias in machine learning models, paving the way for more responsible and transparent AI applications.**

***Index Terms*—Artificial Intelligence, Machine Learning, Fairness, Bias, Causal Models, Grammatical Evolution, Causal Bayesian Networks**

I. Introduction

Unwanted bias refers to the unconscious assumptions that affect our decisions, actions, and interactions [1]. In statistics, bias denotes systematic errors in data collection or analysis that distort results and misrepresent the true characteristics of the population [2]. These biases stem from societal, behavioral, and cognitive factors such as culture, personal experiences, stereotyping, confirmation bias, authority bias, and overestimation bias. They can also arise from involuntary factors, including anchoring, recall bias, recency bias, the halo effect, and attribution bias [3]. While not all biases are harmful, unwanted biases can perpetuate discriminatory behaviours or actions, resulting in unfair treatment and manifesting cycles of inequality and injustice [4].

Artificial Intelligence (AI) systems were built to aid decision-making [5], but unwanted bias can perpetuate social inequalities and discrimination. Mitigation strategies are essential to ensure fair, transparent and unbiased systems. Since an AI system's decisions are dependent on training data [6], biases may emerge in critical systems such as facial recognition system [7], decision support systems in criminal justice systems, healthcare, and financial systems [8]. It is essential to recognize and address these unwanted biases to prevent unintended consequences and promote a more equitable and just society. Protected attributes such as race, gender, or age are legally protected against discrimination [8]. Groups that are at a disadvantage are called *unprivileged*, while those that benefit are referred to as *privileged*. Unprivileged groups often face potential disadvantage compared to privileged groups in contexts like hiring, and financial lending [6]. Causal theory is a fundamental framework for analysing dependencies and understanding causal relationships in complex systems [9].

The goal of our approach is to learn Causal Bayesian Networks (CBNs) from causal structures and identify the most effective causal graph that optimises fairness and accuracy. The rest of the paper is organized as follows. The related work is discussed in Section II, followed by methodology in Section III. We present our results in Section IV which are discussed in Section V. Finally, we offer some conclusions and propose the future directions in Section VI.

II. Related Work

In this section, we review existing work on causal methods and Evolutionary Computation (EC) techniques used in Machine Learning (ML), specifically focusing, causal approaches, and Multi-objective Optimisation (MOO).

A. Bias Mitigation

Bias mitigation in ML is a critical area of research, with various approaches proposed to address the issue. These methods can be broadly categorized into pre-processing, in-processing, and post-processing techniques [10]. Pre-processing techniques aim to modify the training data to reduce bias before model training. In-processing techniques focus on adjusting the learning algorithm itself to ensure fairness during the training process. Post-processing techniques adjust the model's predictions after training to achieve fairness. Our approach uses in-process bias mitigation to achieve our goals.

B. Causal Methods

A causal model enables predictions, causal inference, and reasoning about interventions and counterfactual scenarios [9]. It is formally represented as a Directed Acyclic Graph (DAG), where nodes correspond to variables (X_i), and edges direct causal relationships from parent nodes (P_i), with noise variables as external influences [11]. CBNs are probabilistic graphical models representing the dependency structure of

the variables and edges represent their joint distribution. The dataset can be used to learn and induce a CBN, let us call it B, that encodes a distribution $P_B(A_1, ..., A_n, c)$, from a given training set. Thereby given a set of attributes $a_1, ..., a_n$, the classifier based on B returns the label c that maximises the posterior probability $P_B(c|a_1, ..., a_n)$ [12]. The difference between CBNs and Bayesian Networks (BNs) is that CBNs are used to model causal relationships between variables, while BNs are used to model statistical dependencies between variables [9]. Learning the dependency structure of a DAG is NP-hard [13]. Therefore, we need optimisation methods, such as the Non-combinatorial Optimisation via Trace Exponential and Augmented lagRangian for Structure Learning (NOTEARS) algorithm [14], Peter and Clarke (PC) algorithm, score-based methods, and constraint-based methods to learn the structure of a CBN from data [15]. We use the Python CausalNex Library [16] to create causal structures and CBNs for prediction and interpreting causal relationships.

C. Evolutionary Computation Methods

Grammatical Evolution (GE) algorithm is an EC method that searches the solution space, specified by a grammar, using evolutionary operators based on natural selection [17]. The process starts with an initial population of candidate solutions, encoded as genotypes using integer-based encoding for efficiency [18]. Each genotype is mapped to a phenotype using a grammar, evaluated by a fitness function, and selected by methods like roulette wheel or tournament [19].

A number of previous studies have used EC techniques to create BNs. The Grammar-Guided Evolutionary Bayesian Networks approach uses GE and Evolutionary Algorithm (EA) to automate BN creation. A Context Free Grammar (CFG) ensures syntactically correct structures, while genetic operators iteratively evolve these networks. This method efficiently finds optimal or near-optimal configurations representing the data [20]. The grammar structure is not able to create a wide range of graphs, on account that they focus on the order of variables. Moreover, their focus was not on addressing fairness. A hybrid Evolution-Guided Bayesian Optimization (EGBO) algorithm introduces selection pressure to decrease sampling wastage, which not only determines the Pareto Front efficiently but also achieves better coverage of the Pareto Front while limiting sampling in the infeasible space [21]. The algorithm combines evolutionary strategies with Bayesian optimization to enhance the search process. The mutation process is primarily used to evolve BNs, allowing for diversity and exploration of the solution space. This improves the accuracy and reliability of learned BNs by ensuring that a wide range of potential solutions is considered [22]. However, this method does not focus on fairness and accuracy considerations, which are critical for developing unbiased and equitable AI systems. Our research aims to address this gap by integrating fairness metrics into the evolutionary process, ensuring that the resulting models are both accurate and fair.

D. Multi-Objective Optimisation

A number of previous studies have used EC techniques to address fairness and bias using multi-objective optimisation methods. A multi-objective approach to fairness in ML addresses both disparate treatment and disparate impact [23]. This framework for learning classifiers without disparate mistreatment aims to minimize the overall harm experienced by different groups while maintaining high predictive accuracy. The approach involves incorporating fairness constraints into the learning process. This approach leads to decision boundaries that minimize disparate treatment and impact, thereby reducing the overall harm to disadvantaged groups [24].

Another comprehensive study of multi-objective fairness in ML focuses on optimizing multiple fairness objectives simultaneously. This study introduces a general framework for incorporating fairness constraints into the optimisation process, allowing practitioners to balance fairness with other objectives such as accuracy and efficiency. It provides theoretical insights and practical algorithms for addressing multi-objective fairness in various ML tasks [25]. The authors define the problem of fair regression in terms of a constrained optimisation problem with statistical parity and bounded group loss as constraints. The results are focused on regression and therefore not suitable for classification tasks.

Fairness Oriented Multiobjective Optimisation (FOMO) is a concept that advocates that an ML model's fairness improvement is a task that requires multi-objective optimisation, often with conflicting criteria [26]. The authors emphasize the use of multi-objective fairness trade-offs instead of transforming error-fairness trade-off into a single-objective problem. FOMO defines the fair ML task as one of solving weights classification problem with multiple objectives. Evolutionary Multi-Objective Optimization (EMO) is proposed so that the problem remains tractable [26]. The authors have implemented the system using Non-dominated Sorting Genetic Algorithm II (NSGA-II) [27].

Our focus on learning causal models with fairness and accuracy is different from these approaches. We are creating a set of causal graphs and exploring the solution space using EC. We propose the use of CFG to encode the causal graphs.

III. Methodology

This section describes the details of our proposed method.

TABLE I: PonyGE2 Parameters

Parameter	Value
CROSSOVER_PROBABILITY:	0.75
GENERATIONS	50
GENERATION_SIZE	95
MUTATION	flip per codon
MUTATION_EVENTS	1
POPULATION_SIZE	100

A. Experimental Setup

The experiment was carried out using the PonyGE2 library [28]. The parameters set for the experiment are in Table

TABLE II: German Credit Dataset class distribution

Creditability	Sex	Train	Test	Total
Bad	Male	398	101	499
Bad	Female	162	39	201
Good	Male	146	45	191
Good	Female	93	16	109
	Total	799	201	1000

I. These parameters include the population size, number of generations, mutation events, genome length, codon size, and other important factors. A mutation event of 1 flip per codon was set. The dataset used for the experiments was the German Credit dataset [29]. Since it is a relatively small dataset, the number of generations and individuals were reduced accordingly. The class distribution with respect to sex, for both train and test dataset is shown in Table II. The dataset is not balanced with respect to the class. It contains more instances of males than females and more instances of bad credit than good credit. The overall evolutionary process is detailed in Algorithm 1.

Algorithm 1 Grammatical Evolution algorithm for bias mitigation.

```
Create the population of individual solutions using position independent grow
while termination condition not met do
    for g ∈ population do
        Map the genotype g to create an edge list using the grammar
        Use the edge list to create a DAG
        Use the DAG to learn CBN from training dataset
        make predictions for the test dataset from the CBN
        Evaluate the fitness of each individual
    end for
    Select individuals based on fitness
    Apply crossover and mutation to create offspring
    Replace the least fit individuals in the population with best fit offspring
end while
Output best individual
```

B. Grammar

We utilize the CausalNex library, which is well-suited for constructing CBNs. The GE algorithm employs a grammar to convert a genotype, represented as a sequence of integers, into a phenotype, which is the actual solution. During the mapping phase, each genotype in the population is transformed into a phenotype based on a predefined grammar. This grammar is designed to fit the representation of the possible solutions. In this experiment grammar is defined to facilitate the generation of valid causal graphs to analyse the German credit dataset [29]. The following Backus-Naur Form (BNF) grammar has been used in these experiments.

1) edgelist ::= edges
2) edges ::= edge | ⟨edge, edges⟩
3) edge ::= (feature, feature) | (feature, class)
4) feature ::= un_feature | p_feature
5) un_feature ::= Account Balance | Purpose | ...
6) p_feature ::= Sex | Age
7) class ::= Creditability

The input for the CausalNex function to create a CBN is an edge list of the graph. The grammar is designed to output a causal graph's edge list, which is crucial for this process. Production 1 of the grammar specifies the creation of edges. These edges are the connections between different nodes (features or class) in the causal graph. Production 2 allows for the definition of edges that can be either single edges or recursively defined sets of edges. This flexibility supports the creation of complex graph structures. Production 3 defines that edges can connect a feature to another feature or from a feature to the class. Production 4 specifies that features can be categorized as either unprotected or protected. Unprotected features are listed in production 5 (only two are shown here for brevity), while protected features are listed in production 6. Terminal production 7 explicitly defines the class feature, in this case, "Creditability". This is the target variable in the dataset.

C. Fitness Functions

Fitness functions play a crucial role in directing the EC towards a specified goal. In this context, the edge list representing each graph is used to construct a CBN. Subsequently, classification is performed using the training dataset and the CBN derived from the causal graph. The model's predictions are then used to test the fitness of every individual of the population.

1) Fairness: Let Y be the ground truth or real class labels, $\hat{Y}$ be model's predictions, $A = 0$ is unprivileged and $A = 1$ the privileged values for sensitive or protected attributes. The True Positive Rate (TPR) and False Positive Rate (FPR) are defined as follows. The TPR is defined as the ratio of the True Positives (TPs) to the total number of actual positives, TPs and False Negatives (FNs). The FPR is defined as the ratio of the False Positives (FPs) to the total number of actual negatives, FPs and True Negatives (TNs).

$$TPR = \frac{TP}{TP + FN} \quad FPR = \frac{FP}{FP + TN} \tag{1}$$

Equalized Odds is a fairness metric, where the classifier is considered fair if it provides equal TPR and FPR across different demographics or protected groups [30]. The following equation states the principle more specifically.

$$P(\hat{Y} = 1|Y = y, A = 0) = P(\hat{Y} = 1|Y = y, A = 1), y \in 0, 1 \tag{2}$$

where y is the actual class label, A is the protected attribute, and $\hat{Y}$ is the predicted class label. For $y = 1$, the equation shows that the TPR is equal across both demographics and for $y = 0$, the equation indicates that the FPR is also equal. This means that the classifier achieves fairness by having equal

TPR and FPR for both privileged and unprivileged groups. In other words, the thresholds where both TPR and FPR are calculated are the same across groups, indicating that the classifier is not biased with respect to these performance metrics. In certain scenarios, the outcomes of $Y = 1$ is considered as the "advantageous" such as in cases of loan repayment, college admission, or receiving a promotion. A relaxation of the equalized odds defined in equation 2 is to only require the non-discrimination within the advantaged outcome group $Y = 1$ [30]. The equal opportunity is a relaxed form of equalized odds where the condition is that only the TPR for both demographics is the same.

$$P(\hat{Y} = 1|Y = 1, A = 0) = P(\hat{Y} = 1|Y = 1, A = 1) \quad (3)$$

To prioritize fairness the fitness function in equation 4 is used to evaluate the fitness and select the best individual. This is the Equal Opportunity Difference (EOD) and can be calculated from predictions.

$$\min\{|\text{TPR}(\text{Sex} = \text{Female}) - \text{TPR}(\text{Sex} = \text{Male})|\} \quad (4)$$

EOD is used as the fairness metric in this experiment to evaluate fairness. We will focus on other fairness metrics as part of future work.

2) Accuracy: Accuracy used as fitness function, is a measure of a model's performance, defined as proportion of correct predictions out of the total predictions and can be calculated using equation 5.

$$Accuracy = \frac{TN + TP}{TP + TN + FP + FN} \quad (5)$$

D. Multi-Objective Optimisation

A preliminary version of this work considered these fitness functions independently [31]. We develop the previous work here using multi-objective experiment was conducted using the NSGA-II algorithm [27] implemented in PonyGE2 [28]. The fitness functions used in the MOO are the accuracy and fairness.

$$accuracy\_score = \frac{1}{1 + accuracy} \quad (6)$$

A Pareto front represents a set of non-dominated solutions, where no solution in the set is better than another across all objectives, and each is better than the remaining solutions in the initial population. The Pareto front is found using EOD and the accuracy score defined in equation 6. This allows for the simultaneous minimization of both objectives: EOD and the accuracy score. The best solution among those in the Pareto front was selected using the Hypervolume (HV) metric [32]. The HV of a set of non-dominated points S in two dimensions, with respect to a reference point $r = (r_x, r_y)$, is computed as:

$$HV(S, r) = \sum_{i=1}^{n} (r_x - x_i) \cdot (r_y - y_i)$$

where (x_i, y_i) are the coordinates of the Pareto-optimal solutions, and r_x, r_y are the coordinates of the reference point.

IV. Results

The following results are based on 30 GE experiment runs. We report the best solution, which is the best across all the runs, and average performance (mean across runs). The experiment was conducted using the NSGA-II algorithm [27] implemented in PonyGE2 [28]. The experiment's 30 runs resulted in 498 first fronts. The Pareto Front for fairness vs 1/(1 +accuracy) for all the solutions generated by the runs is shown in Fig. 2. The average and best test values for both the objectives is shown in Table III. The causal graphs of the best individuals using the Pareto Front, are shown in Fig. 1. The test EOD is 0.004167 and test accuracy is 0.687 for the three causal graphs. The average graph height of the Pareto front

TABLE III: Summary of Test Fairness and Test Accuracy Metrics

	Mean	Best
Fairness	0.101064	**0.004167**
Accuracy	0.669207	**0.711443**

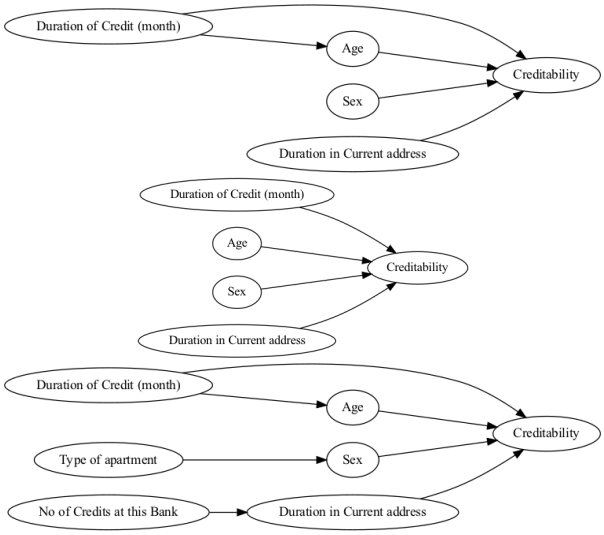

Fig. 1: DAG Best Solutions (EOD : 0.004167, accuracy 0.687).

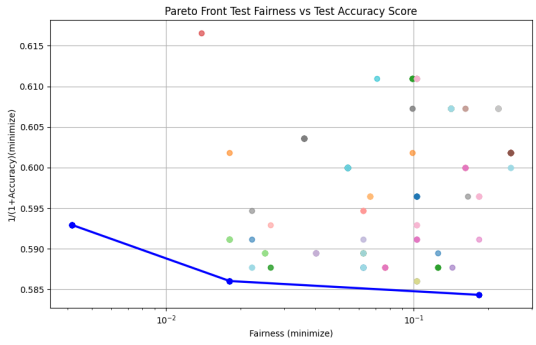

Fig. 2: Pareto fronts test accuracy score vs test fairness

DAGs was 1.80, indicating that the graphs are non-trivial and

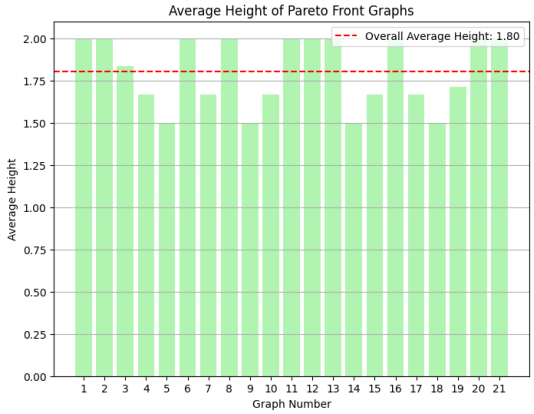

Fig. 3: Average Graph Heights for the Pareto Fronts

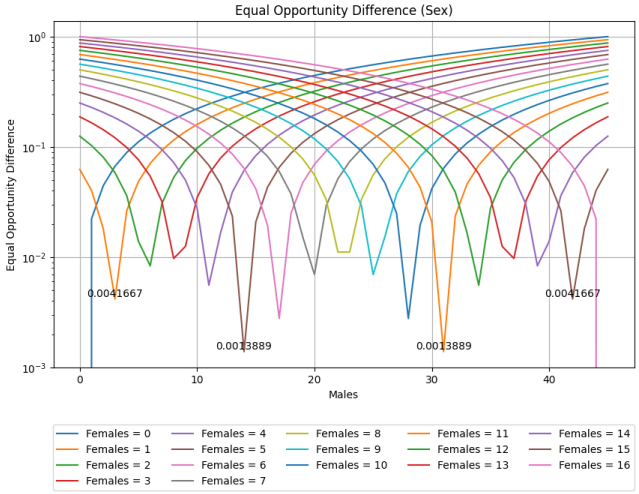

Fig. 4: Equal Opportunity Difference for different values of TPR for test dataset.

capable of capturing complex dependencies among features. The average graph height is illustrated in Fig. 3.

V. Discussion

The MOO experiment showed promising results, since both fitness functions were optimised resulting in Pareto fronts.The graph in Fig. 4 illustrates the values of EOD across different values of predicted positive males, range 0 to 45 shown in x-axis and predicted positive females range 0 to 16 represented by coloured lines. The EOD is computed using equation 4. Apart from the extremal cases where EOD is zero at (0,0), (45 males, 16 females). The minima marked on the graph and is 0.000139 at (14 males, 5 females), (31 males, 11 females). These fairness values were achieved in training. However, the EOD on test dataset as indicated in the Table III and marked on graph is 0.004167, also shown in Fig. 4. The three best DAG based solutions, all achieving the same test fairness and test accuracy are shown in Fig. 1. The flexibility of EC techniques, where multiple solutions can be found that are equally optimal, is important for causal inference, as it allows for the identification of causal relationships between variables. The first two graphs have the same features as nodes; however, the difference lies in how the duration of credit (in months) affects other features. In the first graph, it influences both age and creditability, whereas in the second graph, it affects only creditability. The third graph has the first graph as a sub-graph and two more features are added, type of apartment affects Age, and current credits at this bank affects duration in current address. Causal interventions are needed to confirm if they are causal features, which we will look at next in future work. However, here we believe that these are good candidates for causal features. Given this information, we may be able to interpret the causality of features and make informed decisions about the model depending on the domain. Apart from providing the best solutions, our approach also provides other options such as shown in Fig. 5. The DAG has a test fairness of 0.02639 and test accuracy of 0.7015. This is a good example of the trade-off between fairness and accuracy. The DAG is not optimal but provides a better accuracy than the

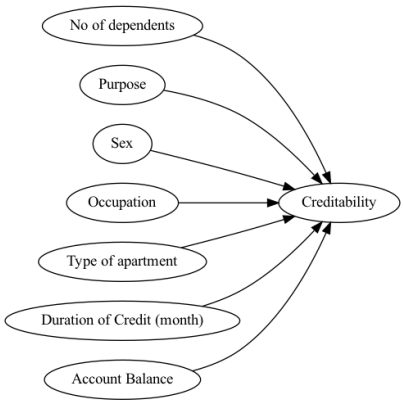

Fig. 5: DAG with Test Fairness 0.02639 and test accuracy 0.7015

best solutions. This is important for practitioners, as it allows them to choose the best solution based on their requirements.

An analysis of the number of features included in the first front BN, as shown in Table IV, reveals that purpose, duration of credit, age, payment status of credit, account balance, and duration at current address are among the most frequently occurring features. This suggests a strong correlation between these features and the target variable

TABLE IV: Direct Connections to Creditability

Feature	Connections
Purpose	17
Duration of Credit(months)	16
Age	5
Payment Status Credit	4
Account Balance	4
Duration in Current Address	3
Sex	2
Telephone	1
No of Credits Bank	2

VI. Conclusions

The goal of this research was to answer two primary questions. The first was whether EC techniques, such as GE, can be used to build BNs from datasets while optimizing for both fairness and accuracy. The second question was whether DAGs could be used to explain the dependency of the target variable on input features. The results in Section IV indicate that the first objective was achieved and further improved through MOO. The second goal was also met, as the constructed DAGs not only enhance the explainability of ML predictions but also enable causal inference such as identifying potential confounders, estimating causal effects and also supporting counterfactual analysis. Future directions of our research include exploration of additional fairness metrics to provide more comprehensive evaluation of fairness. We also plan to extend our experiments to a broader range of datasets. Causal inference techniques could further enhance the performance and interpretability of the CBNs.

Acknowledgment

This research was supported through the HEA's Technological University Transfer Fund (TUTF) and Dundalk Institute of Technology (DkIT).

References

[1] C. FitzGerald and S. Hurst, "Implicit bias in healthcare professionals: a systematic review," *BMC medical ethics*, vol. 18, pp. 1–18, 2017.

[2] S. Farquhar, Y. Gal, and T. Rainforth, "On statistical bias in active learning: How and when to fix it," 2021. [Online]. Available: https://arxiv.org/abs/2101.11665

[3] D. Kahneman, A. Tversky *et al.*, *Intuituve Prediction: Biases and Corrective Procedures*. Decision Research, Perceptronics, 1977.

[4] D. Kahneman, P. Slovic, and A. Tversky, *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press, 1982. [Online]. Available: https://books.google.ie/books?id=_0H8gwj4a1MC

[5] D. Castro and J. New, "The promise of artificial intelligence," *Center for data innovation*, vol. 115, no. 10, pp. 32–35, 2016.

[6] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM computing surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.

[7] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. PMLR, 2018, pp. 77–91.

[8] J. Chen, N. Kallus, X. Mao, G. Svacha, and M. Udell, "Fairness under unawareness: Assessing disparity when protected class is unobserved," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, ser. FAT* '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 339–348. [Online]. Available: https://doi.org/10.1145/3287560.3287594

[9] J. Pearl, *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2009.

[10] T. P. Pagano, J. G. de Souza, D. S. de Oliveira, D. S. V. de Medeiros, J. V. de Souza, and E. S. de Almeida, "Bias and unfairness in machine learning models: A systematic review on datasets, tools, fairness metrics, and identification and mitigation methods," *Big Data and Cognitive Computing*, vol. 7, no. 1, p. 15, 2023. [Online]. Available: https://www.mdpi.com/2504-2289/7/1/15

[11] S. Chiappa, "Path-specific counterfactual fairness," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 7801–7808, Jul. 2019. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/4777

[12] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Machine learning*, vol. 29, pp. 131–163, 1997.

[13] D. M. Chickering, *Learning Bayesian Networks is NP-Complete*. New York, NY: Springer New York, 1996, pp. 121–130. [Online]. Available: https://doi.org/10.1007/978-1-4612-2404-4_12

[14] X. Zheng, B. Aragam, P. K. Ravikumar, and E. P. Xing, "Dags with no tears: Continuous optimization for structure learning," *Advances in neural information processing systems*, vol. 31, 2018.

[15] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

[16] P. Beaumont, B. Horsburgh, P. Pilgerstorfer, A. Droth, R. Oentaryo, S. Ler, H. Nguyen, G. A. Ferreira, Z. Patel, and W. Leong, "CausalNex," Oct. 2021. [Online]. Available: https://github.com/quantumblacklabs/causalnex

[17] M. O'Neill and C. Ryan, "Grammatical evolution," *IEEE Transactions on Evolutionary Computation*, vol. 5, no. 4, pp. 349–358, 2001.

[18] J. Hugosson, E. Hemberg, A. Brabazon, and M. O'Neill, "Genotype representations in grammatical evolution," *Applied Soft Computing*, vol. 10, no. 1, pp. 36–43, 2010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1568494609000611

[19] D. E. Goldberg and K. Deb, "A comparative analysis of selection schemes used in genetic algorithms," in *Foundations of genetic algorithms*. Elsevier, 1991, vol. 1, pp. 69–93.

[20] J. M. Font, D. Manrique, and E. Pascua, "Grammar-guided evolutionary construction of bayesian networks," in *Foundations on Natural and Artificial Computation*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 60–69.

[21] A. Low, F. Mekki-Berrada, A. Gupta, A. Ostudin, J. Xie, E. Vissol-Gaudin, Y.-F. Lim, Q. Li, Y. Ong, S. Khan, and K. Hippalgaonkar, "Evolution-guided bayesian optimization for constrained multi-objective optimization in self-driving labs," *npj Computational Materials*, vol. 10, 05 2024.

[22] Y. Shen, T.-A. Hoang, W. Buntine, and H. K. Dam, "Bayesian network structure learning by incorporating a mutation process into the evolutionary algorithm," in *Proceedings of the 2011 European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*. Springer-Verlag, 2011, pp. 451–466.

[23] M. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi, "Fairness constraints: A flexible approach for fair classification," *Journal of Machine Learning Research*, vol. 20, no. 75, pp. 1–42, 2019.

[24] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi, "Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment," *Proceedings of the 26th International Conference on World Wide Web*, pp. 1171–1180, 2017.

[25] A. Agarwal, M. Dudík, and Z. S. Wu, "Fair regression: Quantitative definitions and reduction-based algorithms," in *International Conference on Machine Learning*. PMLR, 2019, pp. 120–129.

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

[27] K. Deb, S. Agrawal, A. Pratap, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: Nsga-ii," in *International conference on parallel problem solving from nature*. Springer, 2002, pp. 849–858.

[28] M. Fenton, J. McDermott, D. Fagan, S. Forstenlechner, E. Hemberg, and M. O'Neill, "Ponyge2: Grammatical evolution in python," in *Proceedings of the and Evolutionary Computation Conference Companion*, 2017, pp. 1194–1201.

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

[30] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," *CoRR*, vol. abs/1610.02413, 2016. [Online]. Available: http://arxiv.org/abs/1610.02413

[31] Z. Irfan, R. Loughran, M. A. Raja, and F. McCaffery, "Multi-objective fairness approach using causal bayesian networks & grammatical evolution," in *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO '25 Companion)*. Malaga, Spain: Association for Computing Machinery, July 14–18 2025.

[32] A. Auger, J. Bader, D. Brockhoff, and E. Zitzler, "Hypervolume-based multiobjective optimization: Theoretical foundations and practical implications," *Theoretical Computer Science*, vol. 425, pp. 75–103, 2012.