Evaluating Multi-objective Evolutionary Algorithms for Fairness-Aware Decision Tree Tuning

Ana Januário[1(✉)], Maria João Alves[1,2], and Carlos Henggeler Antunes[2,3]

[1] CeBER, Faculty of Economics, University of Coimbra, Coimbra, Portugal
uc2023140299@student.uc.pt, mjalves@fe.uc.pt
[2] Institute of Systems and Computer Engineering (INESC), Coimbra, Portugal
[3] Department of Electrical and Computer Engineering, University of Coimbra, Coimbra, Portugal
ch@deec.uc.pt

Abstract. Fairness is increasingly recognised as a critical requirement in machine learning applications, particularly in sensitive domains such as healthcare, where misclassifications may disproportionately affect specific demographic groups. This study investigates how multi-objective hyperparameter optimisation can balance predictive performance and algorithmic fairness in decision tree models. The optimisation model simultaneously minimises classification error and gender-based unfairness, defined as disparities in true positive rates between male and female groups. Three optimisation strategies are compared: an exhaustive grid search and two evolutionary multi-objective algorithms, NSGA-II and MOEA/D. Their resulting solution sets are analysed by comparing Pareto fronts, receiver operating characteristic (ROC) curves and gender-stratified confusion matrices to assess both overall predictive performance and the distribution of classification errors across demographic groups.

Grid search identifies the configuration with the highest area under the ROC curve due to its exhaustive exploration of the hyperparameter space. NSGA-II significantly reduces computational time while retaining competitive solutions, though with slightly lower discriminative capability. MOEA/D discovers configurations that closely match the predictive performance of grid search and display more balanced error distributions across genders. These results show the potential of evolutionary multi-objective optimisation as an effective and practical approach for fairness-aware hyperparameter tuning. They also highlight the importance of assessing group-specific error distributions beyond aggregate fairness metrics to obtain a more reliable and nuanced understanding of model behaviour in sensitive decision-making contexts.

Keywords: Multi-Objective Optimisation · Hyperparameter Optimisation · Machine Learning

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16769, pp. 274–291, 2027.
https://doi.org/10.1007/978-3-032-30494-0_18

1 Introduction

Decision tree classifiers remain widely used in machine learning due to their intuitive structure, ease of interpretation, and ability to handle heterogeneous data types [4,13]. In their classical formulation, trees recursively partition the feature space into increasingly homogeneous regions with respect to the target variable. At each internal node, a splitting rule is selected to maximise an impurity reduction criterion, while terminal nodes assign class labels based on the majority class of the training samples reaching that node. This hierarchical and rule-based representation makes decision trees particularly attractive in decision support applications, where model transparency and interpretability are essential [2,13].

Despite their conceptual simplicity, the predictive behaviour of decision trees is strongly influenced by a set of hyperparameters that control the tree-growing process. Parameters such as maximum depth, minimum number of samples required to split a node and pruning thresholds determine the complexity of the resulting model and its ability to generalise beyond the training data. Trees grown with weak constraints may capture complex relationships at the risk of overfitting, whereas stricter constraints typically produce simpler models with improved generalisation but reduced flexibility. Consequently, the careful selection of hyperparameter configurations is crucial to achieving an appropriate balance between predictive accuracy, robustness and interpretability [1,13].

Hyperparameter optimisation seeks to identify configurations of the control parameters that maximise model performance according to a chosen evaluation criterion. Traditional approaches, such as manual tuning or exhaustive grid search, evaluate predefined sets of candidate configurations and select the one that optimises a chosen validation metric [1]. Although straightforward to implement, these strategies become computationally expensive as the number of hyperparameters increases and may struggle to capture complex trade-offs when multiple objectives must be considered simultaneously. For this reason, more advanced search strategies based on evolutionary algorithms and other meta-heuristic techniques have been increasingly explored to navigate high-dimensional hyperparameter spaces more efficiently [9,11,15,18].

While most hyperparameter optimisation procedures focus primarily on improving predictive performance, growing attention has been paid to the societal implications of machine learning systems. A substantial body of research has highlighted systematic disparities in model outcomes across demographic groups, bringing algorithmic fairness to the forefront of machine learning research [5,14]. In classification tasks, unfairness may arise when models display different error rates, true positive rates or decision thresholds for protected groups defined by attributes such as gender, race or age [12]. Fairness-aware learning, therefore, seeks to design models that satisfy explicit fairness criteria while maintaining acceptable predictive performance [14,17]. Bias mitigation can be addressed at different stages of the modelling pipeline: through pre-processing techniques that modify the training data, by incorporating fairness constraints directly into the learning algorithm or by post-processing model outputs to reduce disparities [14].

Among the various fairness concepts proposed in the literature, group-based criteria such as equal opportunity, which requires similar true positive rates across protected groups, are particularly relevant in sensitive domains such as healthcare and credit scoring [12].

Recent work highlights the challenges of optimising fairness in practice. The choice of fairness metric is itself non-trivial: Uddin et al. [17] show that the same dataset can yield fair outcomes under one algorithm and unfair outcomes under another, depending on the fairness definition adopted. Corbett-Davies et al. [5] demonstrate that widely used formal fairness criteria often lead to Pareto-dominated decision policies, suggesting that no axiomatic definition of fairness is universally adequate. d'Aloisio et al. [7] show that, even when fairness is explicitly considered as an objective to be optimised together with predictive performance, the resulting trade-offs depend heavily on the choice of the metric and model type. These findings suggest that, rather than enforcing fairness as a post-hoc constraint, it may be more effective to incorporate it directly into the model selection process through hyperparameter optimisation.

These developments motivate treating hyperparameter selection as a multi-objective optimisation problem, in which predictive performance and fairness are pursued simultaneously rather than independently. Instead of collapsing multiple evaluation criteria into a single scalar objective using ad hoc weights, multi-objective optimisation explicitly searches for Pareto-optimal solutions that represent efficient trade-offs between competing goals [8]. Evolutionary multi-objective algorithms, such as NSGA-II and MOEA/D, are especially well suited to this setting. By maintaining a diverse population of candidate solutions, these algorithms can approximate the Pareto front in a single optimisation run, thereby providing information on different trade-offs between objectives [9,18]. Moreover, both algorithms have demonstrated strong empirical performance across a wide range of multi-objective optimisation problems and have become standard benchmarks in the evolutionary optimisation literature. When applied to decision tree hyperparameter optimisation, these algorithms enable practitioners to explore how different configurations balance classification error against group-level fairness measures, such as disparities in true positive rates between gender groups, and to select models that best align with the requirements of a given decision-making context [6,15].

In this work, we investigate the multi-objective hyperparameter optimisation of decision tree classifiers, where predictive performance and fairness are treated as objectives to optimise. Specifically, we jointly minimise classification error and gender-based unfairness, which is measured as disparities in true positive rates between male and female groups according to the equal opportunity criterion [12]. To explore the trade-offs between these objectives, we compare an exhaustive grid search strategy with two evolutionary multi-objective algorithms, NSGA-II and MOEA/D, and analyse the resulting Pareto fronts obtained under each optimisation approach [9,15,18]. The proposed framework aims to provide a systematic approach for examining how different hyperparameter configura-

tions affect the trade-off between predictive accuracy and group-level fairness in decision tree models.

The main contributions of this work are:

- a multi-objective framework for fairness-aware hyperparameter tuning;
- a comparative analysis of grid search and well-established evolutionary optimisation methods;
- a detailed group-level error analysis using confusion matrices.

2 Methodology

In the framework of a decision tree classifier for predicting heart disease, this study develops a model selection structure formulated as a multi-objective hyperparameter optimisation problem. Different decision tree hyperparameter configurations are generated and explored using grid search and evolutionary search strategies to optimise predictive performance and fairness with respect to gender.

To ensure rigorous evaluation, the dataset was randomly partitioned into three mutually exclusive subsets: a training set used for model fitting, a validation set for identification of Pareto-optimal solutions for hyperparameter tuning and an independent test set for final performance assessment.

Hyperparameter optimisation was carried out using three distinct approaches: exhaustive Grid Search, the Non-dominated Sorting Genetic Algorithm II (NSGA-II) [8,9] and the Multi-Objective Evolutionary Algorithm based on Decomposition (MOEA/D) [18], enabling a comparative analysis of their efficiency and solution quality.

The hyperparameter search space has been designed to explore the trade-off between model complexity and predictive performance. It includes the following parameters:

- Maximum tree depth (max_depth): integer values from 1 to 20.
- Minimum number of samples required to split a node (min_samples_split): integer values from 2 to 50.
- Maximum number of features considered for each node split (max_features): integer values from 1 to 7.

2.1 Performance Metrics

Model predictions were used to compute several performance metrics. Let TP, FP, and FN denote the numbers of true positives, false positives, and false negatives, respectively.

Precision and recall are defined in (1) and (2):

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$TPR(Recall) = \frac{TP}{TP + FN} \tag{2}$$

Recall is also referred to as the true positive rate (TPR), as it measures the proportion of actual positive instances that are correctly identified by the model. These two measures are combined in the $F1_{score}$ ($F1$), which corresponds to the harmonic mean of precision and recall, as in (3):

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{3}$$

The classification error (E) considered in this study is defined in (4):

$$E = 1 - F1 \tag{4}$$

To evaluate fairness considering groups, the recall was computed separately for male and female individuals by partitioning the validation and test set according to the `Sex` attribute, thereby enabling the assessment of group-specific diagnostic performance conditional on the presence of heart disease. The unfairness metric (U) was defined as the absolute difference between the true positive rates of the two groups, as in (5):

$$U = |TPR_{men} - TPR_{women}| \tag{5}$$

where TPR_{men} and TPR_{women} denote the recall scores for male and female individuals, respectively.

2.2 Multi-objective Framework

Let $\mathcal{M}$ be the set of all trained models. Model selection was based on Pareto dominance, considering classification error (E) and unfairness (U) as the two objectives to be minimised. A model $M_i \in \mathcal{M}$ is said to dominate another model $M_j \in \mathcal{M}$ if it is not worse than M_j with respect to both objectives and strictly better in at least one of them, that is, with at least one strict inequality (6).

$$E_i \leq E_j \text{ and } U_i \leq U_j \tag{6}$$

The goal is, therefore, to identify the Pareto front, containing the models that offer the best possible trade-offs, where any further reduction in unfairness would inevitably result in a higher classification error and vice versa. In case of ties, where multiple configurations yielded identical error and unfairness values, the Occam's Razor principle was applied, giving preference to the least complex model [2]. Each solution represents a specific decision tree hyperparameter configuration, trained on the training set and evaluated on the validation set to obtain its classification error and unfairness values.

To assess the generalisation performance of the Pareto-optimal solutions, all models identified as Pareto optimal during the validation phase were evaluated

on an independent test set. We further rank Pareto-optimal models using the Area Under the ROC Curve (AUC), which is not an optimisation objective but provides an additional discriminative evaluation. ROC curves were generated for the five models exhibiting the highest AUC values. ROC curves illustrate the relationship between the true positive rate and the false positive rate across all possible decision thresholds, providing a threshold-independent evaluation of model performance. The AUC summarises this relationship into a single scalar value that reflects the model's overall ability to discriminate between positive and negative instances. Therefore, these indicators provide a comprehensive assessment of discriminative performance and support the final model selection [10].

The solutions produced by NSGA-II and MOEA/D were compared to the non-dominated solutions obtained with the exhaustive grid search. For selected models from all optimisation strategies, confusion matrices disaggregated by gender were analysed to examine error distributions across demographic groups and verify fairness properties in greater detail.

2.3 Dataset and Implementation Options

The analysis employed a dataset about personal health indicators associated with heart disease: heart_2020_cleaned dataset, publicly available on the Kaggle platform[1]. It includes a binary target variable, `Heart Disease`, denoting the presence or absence of heart disease, and a set of explanatory variables encompassing demographic, behavioural, and health-related attributes (e.g., `BMI` (body mass index), `Smoking`, `Alcohol Drinking`, `Stroke`, `Physical Health`, `Mental Health`, `Difficulty Walking`, `Sex`, `Age Category`, `Race`, `Diabetic`, `Physical Activity`, `General Health`, `Sleep Time`, `Asthma`, `Kidney Disease`, and `Skin Cancer`).

The target variable is imbalanced: approximately 91.5% of instances correspond to individuals without heart disease, whereas 8.5% represent diagnosed cases. Consequently, accuracy was deemed unsuitable as it may mask poor predictive performance for the minority class. Instead, the $F1_{score}$ was adopted as the primary performance metric, given its robustness to class imbalance and its formulation as the harmonic mean of precision and recall [2,10].

For grid search, a decision tree model was trained for each combination of hyperparameters (max_depth, min_samples_split, max_features), resulting in 6,860 distinct models being fitted on the training set. Each trained model was subsequently evaluated on the validation set.

The evolutionary multi-objective optimisation approaches were implemented using the same decision space. Both NSGA-II [9] and MOEA/D [18] were configured with identical parameter settings to enable a fair comparison. The initial population was generated using an integer random sampling scheme. Variation operators consisted of simulated binary crossover (SBX) with probability

[1] https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease.

$p_c = 0.9$ and distribution index $\eta_c = 15$, and polynomial mutation with probability $p_m = 1/3 \approx 0.33$ and distribution index $\eta_m = 20$. All offspring solutions were repaired using a rounding operator to ensure integer feasibility.

For NSGA-II, the population size was set to $N = 50$, and the algorithm was executed for 60 generations, using non-dominated sorting and crowding distance to promote convergence towards and diversity along the Pareto front [9]. Further experiments with a larger number of generations produced similar results, indicating that the chosen configuration was sufficient to reach stable solutions.

The same termination criterion (60 generations) and population size ($N = 50$) were adopted for MOEA/D. In this case, the algorithm decomposes the bi-objective problem into a set of scalar subproblems defined by uniformly distributed reference directions and employees neighbourhood-based mating. The specific parameters were set to $T = 15$ neighbours and probability of selecting parents from the neighbourhood $p_{\text{neigh}} = 0.7$.

All analyses were conducted in a Python environment. The optimisation algorithms were implemented using the pymoo framework [3] and the decision tree classifier was implemented using the scikit-learn library[2].

3 Results

The exhaustive grid-search procedure evaluated a total of 6,860 candidate models, requiring approximately 20 min of training time. In contrast, NSGA-II evaluated 3,000 candidate solutions in approximately 6 min and 17 s, while MOEA/D assessed 3,060 candidate solutions in 19 min and 6 s.

The Pareto front obtained from the grid search on the validation set includes 15 nondominated solutions, revealing a clear trade-off between predictive performance and fairness (Fig. 1). In this dataset, models with lower classification error tend to exhibit higher unfairness, whereas solutions achieving near-zero unfairness require sacrificing predictive accuracy. Table 1 reports the Pareto-optimal configurations and their objective values. For instance, model G1 attains the lowest error (0.7675) but the highest unfairness (0.0285), while model G15 achieves perfect unfairness at the cost of substantially higher classification error (0.8882).

NSGA-II also produced 15 non-dominated solutions, recovering 9 of the 15 Pareto-optimal configurations identified by the exhaustive grid search (60%). MOEA/D identified 10 Pareto-optimal solutions, recovering 5 of the Pareto-optimal configurations obtained by grid search (50%). Although MOEA/D generated fewer Pareto-optimal solutions, several of its configurations lie close to the Pareto front of the grid search, indicating competitive performance in approximating the trade-off surface. Figure 2 shows the fronts generated by NSGA-II and MOEA/D on the validation set.

Table 2, compares validation and test performance for the grid-search Pareto-optimal configurations. The classification error on the test set remains relatively stable across datasets, ranging from 0.679 to 0.728. In several cases, the test

[2] https://scikit-learn.org.

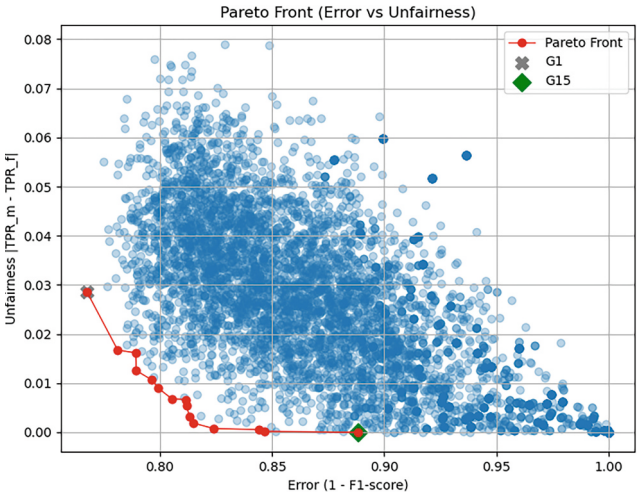

Fig. 1. Pareto front obtained on the validation set, illustrating the trade-off between classification error and gender-based unfairness across decision tree hyperparameter configurations.

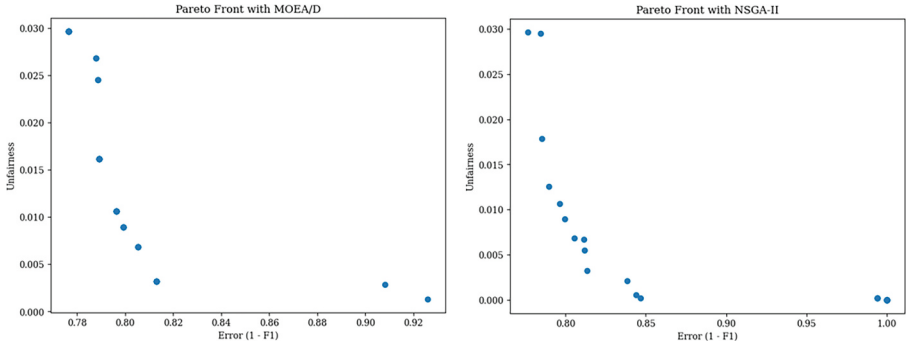

Fig. 2. Pareto front obtained on the validation set for MOEA/D and NSGA-II, illustrating the trade-off between classification error and unfairness across decision tree hyperparameter configurations.

error is even lower than the corresponding validation error, suggesting that the models do not suffer from severe overfitting and display strong generalisation capability.

In contrast, fairness metrics exhibit a markedly different behaviour. Although the validation set unfairness ranges from 0.000 to 0.029, test set unfairness increases substantially, typically falling between 0.09 and 0.20. Even configurations that achieve near-zero unfairness during validation display noticeable disparities on unseen data. This pattern indicates that fairness metrics may be considerably more sensitive to distributional differences between datasets than

Table 1. Pareto-optimal solutions obtained via Grid Search (validation set).

Model	Error	Unfairness	Depth	min_samples_split	max_features
G1	0.7675	0.0285	20	6	7
G2	0.7812	0.0167	19	7	7
G3	0.7893	0.0162	19	2	7
G4	0.7894	0.0125	18	4	6
G5	0.7965	0.0106	19	2	3
G6	0.7994	0.0089	19	2	4
G7	0.8052	0.0068	20	2	2
G8	0.8113	0.0066	17	7	6
G9	0.8119	0.0055	20	4	2
G10	0.8131	0.0032	20	2	1
G11	0.8147	0.0019	15	7	6
G12	0.8241	0.0007	13	15	6
G13	0.8440	0.0005	19	10	1
G14	0.8467	0.0002	19	5	1
G15	0.8882	0.0000	10	9	5

conventional predictive metrics, implying that fairness-optimised configurations may not generalise reliably.

A similar pattern is observed for NSGA-II. As shown in Table 3, test set error remains within a narrow range (0.689 to 0.797), while fairness deteriorates substantially when evaluated out-of-sample. Test set unfairness commonly lies between 0.10 and 0.20, confirming that fairness estimates obtained during optimisation may underestimate true disparities.

The results obtained with MOEA/D follow the same general trend (Table 4). Test set classification error remains consistent with results from grid search and NSGA-II (0.696 to 0.728), whereas fairness again degrades when evaluated out-of-sample. Although validation unfairness is generally below 0.03, the corresponding test values increase substantially for most configurations. Most MOEA/D configurations show test set unfairness above 0.15, suggesting that fairness generalisation issues are persistent across optimisation strategies. Overall, the MOEA/D solutions exhibit comparable predictive performance and fairness behaviour to those obtained with NSGA-II and the grid search.

3.1 Analysis Based on Area Under the ROC Curve

To further identify the most suitable model among the Pareto-optimal solutions, we evaluated the Area Under the ROC Curve on the test set. AUC complements the bi-objective error and fairness evaluation by providing a threshold-independent measure of discriminative ability. Figure 3 presents the ROC curves

Table 2. Results comparison: validation vs. test sets for Pareto-optimal models obtained by the grid-search.

Error (val)	Error (test)	Unfairness (val)	Unfairness (test)	Depth	Samples	Features
0.8241	0.6792	0.0008	0.1436	13	15	6
0.8882	0.6799	0.0000	0.0948	10	9	5
0.8148	0.6855	0.0019	0.1552	15	7	6
0.8114	0.6894	0.0067	0.1594	17	7	6
0.7894	0.6986	0.0126	0.1832	18	4	6
0.7994	0.6999	0.0089	0.1722	19	2	4
0.8440	0.7000	0.0006	0.1481	19	10	1
0.7675	0.7038	0.0285	0.1778	20	6	7
0.7893	0.7053	0.0162	0.1739	19	2	7
0.7812	0.7058	0.0167	0.1781	19	7	7
0.7965	0.7066	0.0107	0.1690	19	2	3
0.8052	0.7176	0.0068	0.1971	20	2	2
0.8467	0.7189	0.0002	0.1484	19	5	1
0.8119	0.7234	0.0055	0.1672	20	4	2
0.8132	0.7281	0.0032	0.1951	20	2	1

Table 3. Results comparison: validation vs. test sets for NSGA-II

Model	Error (val)	Error (test)	Unfairness (val)	Unfairness (test)	Depth	Samples	Features
N1	0.8114	0.6894	0.0066	0.1594	17	7	6
N2	0.7843	0.6962	0.0295	0.1697	20	13	6
N3	0.7894	0.6986	0.0125	0.1832	18	4	6
N4	0.7994	0.6999	0.0089	0.1722	19	2	4
N5	0.8440	0.7000	0.0006	0.1481	19	10	1
N6	0.7854	0.7004	0.0179	0.1749	20	9	6
N7	0.7765	0.7065	0.0297	0.1732	20	2	5
N8	0.7964	0.7066	0.0106	0.1689	19	2	3
N9	0.8382	0.7098	0.0021	0.1284	19	7	2
N10	0.8052	0.7176	0.0068	0.1971	20	2	2
N11	0.8467	0.7189	0.0002	0.1484	19	5	1
N12	0.8119	0.7234	0.0055	0.1672	20	4	2
N13	0.8131	0.7281	0.0032	0.1951	20	2	1
N14	0.9941	0.7603	0.0002	0.0293	3	9	3
N15	1.0000	0.7975	0.0000	0.1006	1	17	1

Table 4. Results comparison: validation vs. test sets for MOEA/D

Model	Error (val)	Error (test)	Unfairness (val)	Unfairness (test)	Depth	Samples	Max features
M1	0.926	0.696	0.001	0.038	8	9	3
M2	0.799	0.700	0.009	0.172	19	2	4
M3	0.788	0.700	0.027	0.183	19	2	5
M4	0.789	0.702	0.025	0.159	20	5	5
M5	0.908	0.705	0.003	0.088	8	9	4
M6	0.789	0.705	0.016	0.174	19	2	7
M7	0.777	0.707	0.030	0.173	20	2	5
M8	0.796	0.707	0.011	0.169	19	2	3
M9	0.805	0.718	0.007	0.197	20	2	2
M10	0.813	0.728	0.003	0.195	20	2	1

for the five Pareto-optimal grid-search configurations with the highest AUC values, and Table 5 summarises their performance.

Table 5. Top 5 Pareto-optimal models obtained by the grid-search ranked by global AUC (test set)

Rank	Depth	Samples	Features	AUC	Error	Unfairness
1	10	9	5	0.8116	0.6799	0.0948
2	13	15	6	0.7877	0.6792	0.1436
3	15	7	6	0.7505	0.6855	0.1552
4	19	10	1	0.7239	0.7000	0.1481
5	17	7	6	0.7129	0.6894	0.1594

The best-performing model in the test set under grid search (depth = 10, min_samples_split = 9, and max_features = 5) achieved the highest AUC (0.8116), a low classification error (0.6799) and the lowest unfairness (0.0948) among the top five models.

A similar AUC-based evaluation was performed for NSGA-II in the test set (Fig. 4). The best configuration achieved an AUC of 0.757, corresponding to a relatively shallow decision tree (depth = 3, min_samples_split = 9, max_features = 3). Although this model exhibits a moderate level of discriminative ability, its AUC remains lower than that obtained by the model identified through grid search. This result suggests that the configuration found by this method prioritised solutions with lower unfairness rather than maximising the AUC.

For MOEA/D, the best configuration achieved an AUC of 0.8086 (depth = 8, min_samples_split = 9, and max_features = 3), which is very close to the best AUC obtained with grid search (0.8116), Fig. 5. Although MOEA/D required more computation time than NSGA II, and even approached that of

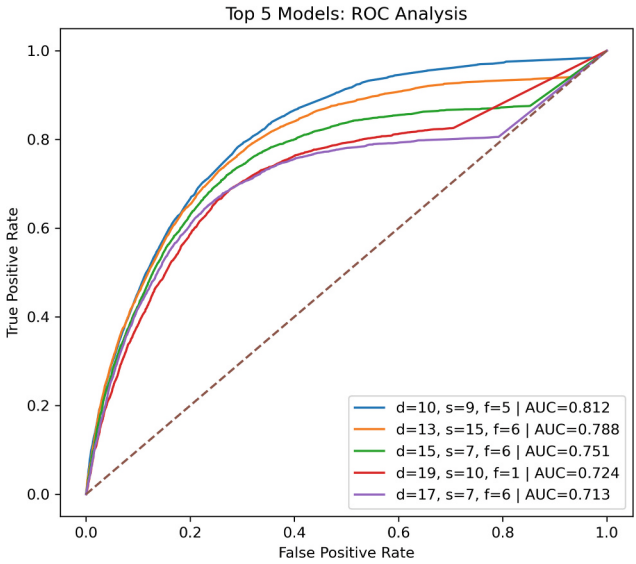

Fig. 3. ROC curves of top-performing models sorted by AUC in the test set for grid search.

grid-search, its top-performing solution demonstrates that decomposition-based evolutionary optimisation can effectively locate high-quality regions of the Pareto front. Unlike NSGA-II, which relies on global non-dominated sorting and crowding distance mechanisms, MOEA/D decomposes the multi-objective problem into multiple scalar subproblems and performs neighbourhood- based updates. The management of these neighbourhood structures may introduce additional computational overhead during the optimisation process.

An interesting pattern was observed when studying the non-dominance considering the classification error and unfairness in the test set. For the solutions obtained with grid search and MOEA/D, we noticed that non-dominated solutions were also the configurations corresponding to the highest AUC values, thus suggesting an alignment between Pareto-optimality in error/unfairness space and overall discriminative performance assessed by AUC. For NSGA-II, three of the five solutions selected for AUC evaluation were non-dominated, including the configuration that achieved the highest AUC.

When examining all non-dominated solutions jointly across methods in the test set, the configuration with the lowest classification error was obtained via grid search, whereas the solution with the lowest unfairness was obtained with NSGA-II (Fig. 6). The solutions obtained by MOEA/D, however, tend to occupy intermediate regions of the Pareto front, suggesting a more balanced trade-off between predictive performance and fairness. Such patterns may also reflect the complex relationship between predictive performance and fairness, where the extent of the trade-off can vary depending on dataset characteristics and on the specific fairness metric considered.

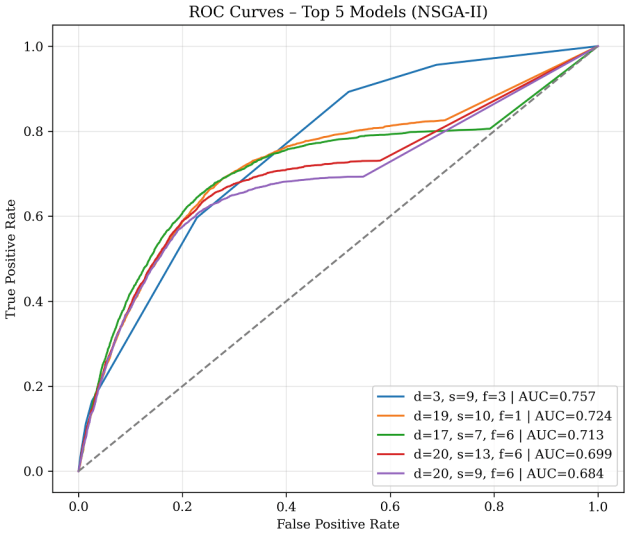

Fig. 4. ROC curves of top-performing models sorted by AUC in the test set during NSGA-II.

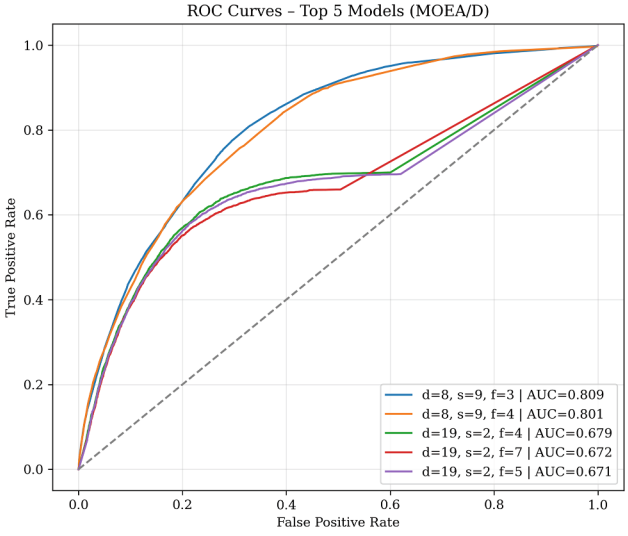

Fig. 5. ROC curves of top-performing models sorted by AUC in the test set during MOEA/D.

3.2 Confusion-Matrix Analysis

While the aggregate unfairness metric provides an overall indication of disparity between demographic groups, it does not reveal how prediction errors are distributed across these groups. The group level confusion-matrix analysis provides

further insights into fairness behaviour. In particular, the False Positive Rate (FPR) is informative because it measures how frequently individuals without the condition are incorrectly classified as positive within each group. Differences in FPR may indicate unequal exposure to unnecessary diagnostic procedures, making this metric especially relevant in clinical decision-support settings.

The FPR is defined in (7):

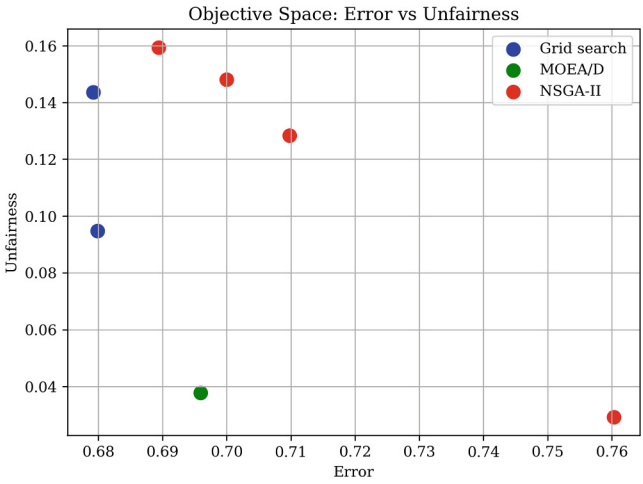

Fig. 6. Non-dominated solutions of each method in the test set.

$$FPR = \frac{FP}{FP + TN} \tag{7}$$

For the best grid search configuration (Fig. 7), recall for men (0.8212) exceeds that for women (0.7264), indicating women are more likely to receive false negatives. Conversely, men exhibit a higher false positive rate (0.3368 vs. 0.2500), implying greater exposure to unnecessary follow-up procedures.

For the best NSGA-II configuration (Fig. 8), recall is substantially higher for both groups (0.9048 for men and 0.8755 for women), but this improvement is accompanied by a significant increase in false positive rates (0.4848 for men and 0.5509 for women). This indicates that the model tends to classify a large fraction of observations as positive, increasing sensitivity at the expense of predictive reliability.

The best MOEA/D configuration (Fig. 9) produces a more balanced distribution of errors: recall values remain competitive (0.8244 for men, 0.7865 for women), and false positive rates are nearly identical across groups (0.3286 for men and 0.3285 for women). This suggests that MOEA/D yields a more symmetric error distribution while maintaining strong discriminative performance.

Overall, the group-level analysis highlights important differences in how each optimisation strategy balances accuracy and fairness. Grid-search achieves the

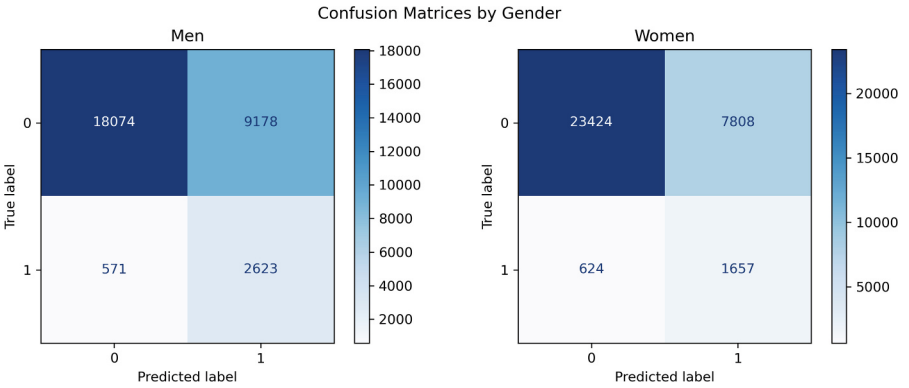

Fig. 7. Confusion matrix stratified by gender (grid search).

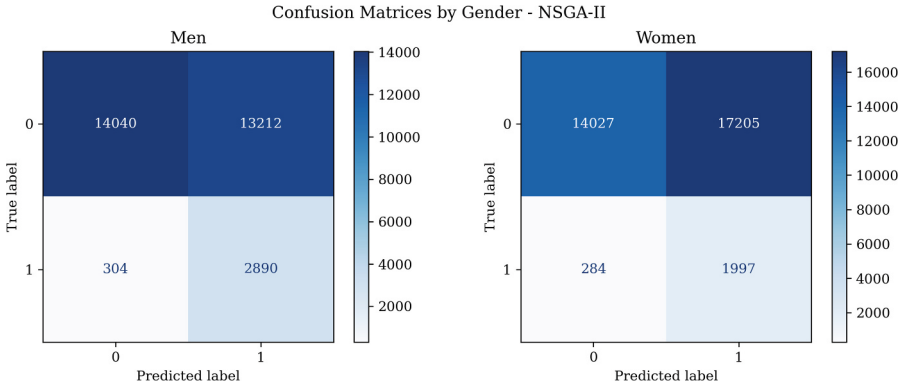

Fig. 8. Confusion matrix stratified by gender (NSGA-II).

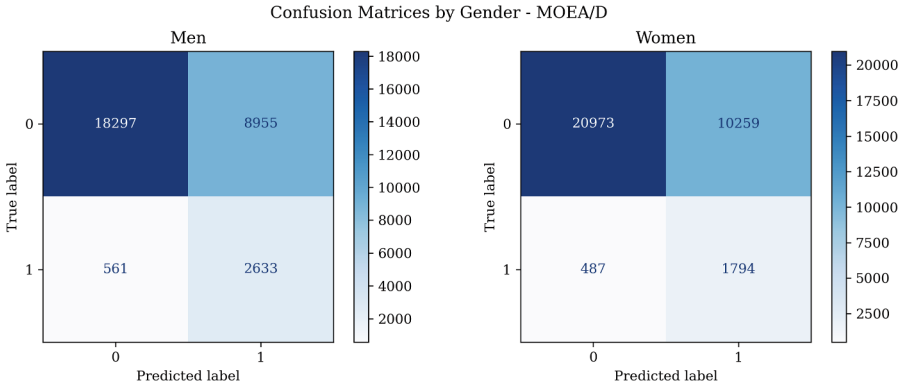

Fig. 9. Confusion matrix stratified by gender (MOEA/D).

strongest discriminative performance, but presents higher sensitivity for men than for women. NSGA-II is the most computationally efficient, but it produces more extreme classification patterns. MOEA/D achieves competitive predictive performance with more balanced error distribution across demographic groups.

4 Conclusions and Future Work

This study investigated the use of multi-objective hyperparameter optimisation to balance predictive performance and fairness in decision tree models. The optimisation problem was formulated with two objectives: minimising classification error and gender-based unfairness, which is measured as disparities in true positive rates under the Equal Opportunity criterion. To analyse the resulting trade-offs, an exhaustive grid search was compared with two evolutionary multi-objective algorithms, NSGA-II and MOEA/D.

The results show that grid search identified the best configuration overall according to the AUC value. However, this exhaustive approach required the greatest computational effort, highlighting its limitations as the hyperparameter space expands. Among the evolutionary approaches, NSGA-II proved to be the most computationally efficient, significantly reducing runtime while still identifying competitive solutions. Nevertheless, its best configurations demonstrated lower discriminative performance (i.e., lower AUC values) compared to those obtained through grid search. In contrast, MOEA/D was able to discover configurations with predictive performance very close to the grid-search optimum, demonstrating its ability to approximate high-quality regions of the Pareto front. Although its runtime was comparable to that of grid search, the resulting models exhibited more balanced false-positive rates across genders, suggesting favourable behaviour in terms of fairness.

The analysis of gender-stratified confusion matrices further revealed that aggregate fairness metrics may obscure important group-level patterns. Even when global unfairness scores appeared small, disparities in recall and false-positive rates remained noticeable. These findings highlight the importance of complementing aggregate fairness measures with detailed group-specific error analysis.

The contribution of this study is to propose a structured multi-objective evaluation framework for fairness-aware hyperparameter tuning. In particular, the comparative analysis of grid search, NSGA-II and MOEA/D highlights their distinct behaviours in terms of fairness–performance trade-offs. The results suggest that evolutionary multi-objective optimisation constitutes a promising alternative to exhaustive search for fairness-aware hyperparameter tuning, especially when computational constraints or large search spaces are present.

This study focuses on decision tree models and a single healthcare dataset. Although being a limitation, this choice allowed for controlled analysis. Future work will extend the analysis to other model families (e.g., ensemble methods and neural networks) with more complex hyperparameter spaces, as well as additional datasets, to assess the robustness and generalisability of the observed patterns.

Fairness was evaluated using the equal opportunity criterion (TPR difference), but other definitions, such as demographic parity and equalised odds, capture different ethical trade-offs and may lead to different model selections. Their incorporation into the multi-objective framework remains an important direction for future research.

Acknowledgements. This work was supported by the Portuguese Foundation for Science and Technology (FCT) under the project grants UIDB/05037/2025 (DOI: 10.54499/UID/05037/2025) and UID/00308/2025 (DOI: 10.54499/UID/00308/2025).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bergstra, J., Bengio, Y.: Random search for hyper-parameter optimization. J. Mach. Learn. Res. **13**, 281–305 (2012). https://dl.acm.org/doi/10.5555/2188385.2188395
2. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, New York (2006)
3. Blank, J., Deb, K.: Pymoo: multi-objective optimization in Python. IEEE Access **8**, 89497–89509 (2020). https://doi.org/10.1109/ACCESS.2020.2990567
4. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and Regression Trees. Wadsworth International Group, Belmont (1984)
5. Corbett-Davies, S., Gaebler, J., Nilforoshan, H., Shroff, R., Goel, S.: The measure and mismeasure of fairness. J. Mach. Learn. Res. **24**(312), 1–117 (2023). http://jmlr.org/papers/v24/22-1511.html
6. Cruz, A., Lawrynowicz, A., Bizarro, P.: Promoting fairness through hyperparameter optimization. In: 2021 IEEE International Conference on Data Mining (ICDM), pp. 114–123. IEEE (2021)
7. d'Alosio, G., Hort, M., Moussa, R., Sarro, F.: FairRF: multi-objective search for single and intersectional software fairness. In: Proceedings of the 2026 IEEE/ACM 48th International Conference on Software Engineering (ICSE-SEIS '26), Rio de Janeiro, Brazil, 12–18 April 2026. ACM, New York (2026)
8. Deb, K.: Multi-objective Optimization Using Evolutionary Algorithms. Wiley, Chichester (2001)
9. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.: A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Trans. Evol. Comput. **6**(2), 182–197 (2002). https://doi.org/10.1109/4235.996017
10. Fawcett, T.: An introduction to ROC analysis. Pattern Recogn. Lett. **27**(8), 861–874 (2006)
11. Feurer, M., Hutter, F.: Hyperparameter optimization. In: Hutter, F., Kotthoff, L., Vanschoren, J. (eds.) Automated Machine Learning. TSSCML, pp. 3–33. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-05318-5_1
12. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. In: Lee, D.D., Sugiyama, M., Luxburg, U.V., Guyon, I., Garnett, R. (eds.) Advances in Neural Information Processing Systems (NIPS 2016), vol. 29, pp. 3315–3323. Curran Associates, Red Hook (2016). https://doi.org/10.48550/arXiv.1610.02413

13. The Elements of Statistical Learning. SSS. Springer, New York (2009). https://doi.org/10.1007/978-0-387-84858-7
14. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Comput. Surv. **54**(6), 115:1–115:35 (2021). https://doi.org/10.1145/3457607
15. Morales-Hernández, A., Van Nieuwenhuyse, I., Rojas Gonzalez, S.: A survey on multi-objective hyperparameter optimization algorithms for Machine Learning. arXiv preprint arXiv:2111.13755 (2021)
16. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo (1993)
17. Uddin, S., Lu, H., Rahman, A., Gao, J.: A novel approach for assessing fairness in deployed machine learning algorithms. Sci. Rep. **14**, 17753 (2024). https://doi.org/10.1038/s41598-024-68651-w
18. Zhang, Q., Li, H.: MOEA/D: a multiobjective evolutionary algorithm based on decomposition. IEEE Trans. Evol. Comput. **11**(6), 712–731 (2007). https://doi.org/10.1109/TEVC.2007.892759