Machine Learning (2024) 113:6539–6559
https://doi.org/10.1007/s10994-024-06583-2

FairMOE: counterfactually-fair mixture of experts with levels of interpretability

Joe Germino[1] · Nuno Moniz[1] · Nitesh V. Chawla[1]

Received: 27 March 2024 / Revised: 10 May 2024 / Accepted: 11 June 2024 /
Published online: 8 July 2024

Abstract

With the rise of artificial intelligence in our everyday lives, the need for human interpretation of machine learning models' predictions emerges as a critical issue. Generally, interpretability is viewed as a binary notion with a performance trade-off. Either a model is fully-interpretable but lacks the ability to capture more complex patterns in the data, or it is a black box. In this paper, we argue that this view is severely limiting and that instead interpretability should be viewed as a continuous domain-informed concept. We leverage the well-known Mixture of Experts architecture with user-defined limits on non-interpretability. We extend this idea with a counterfactual fairness module to ensure the selection of consistently *fair* experts: **FairMOE**. We perform an extensive experimental evaluation with fairness-related data sets and compare our proposal against state-of-the-art methods. Our results demonstrate that FairMOE is competitive with the leading fairness-aware algorithms in both fairness and predictive measures while providing more consistent performance, competitive scalability, and, most importantly, greater interpretability.

Keywords Interpretability · Mixture of experts · Counterfactual fairness · Scalability

1 Introduction

Explainable AI (XAI) has been studied for over three decades (Chandrasekaran et al., 1989), with the objective of providing explanations for learning models' outcomes such that it (1) guarantees the highest level possible of model accuracy, and (2) that human

Editors: Ana Carolina Lorena, Albert Bifet, Rita P. Ribeiro.

✉ Nitesh V. Chawla
nchawla@nd.edu

Joe Germino
jgermino@nd.edu

Nuno Moniz
nuno.moniz@nd.edu

[1] Lucy Family Institute for Data and Society, University of Notre Dame, Notre Dame, IN 46656, USA

actors can understand (Arrieta et al., 2020). Today, explainability usually divides efforts into two groups (Arrieta et al., 2020). On the one hand, post-hoc explanations extract information on more complex (and non-interpretable) models' behavior where the relation between inputs and outputs is complex for humans to understand. On the other hand, interpretable models provide a more transparent view of how model decisions are carried out.

There are several real-world examples of how non-interpretable models deployed in high-risk decision-making environments may incur costly mistakes (Rudin, 2019). In addition, post-hoc explainability cannot fully approximate a black box model and often results in biased and unfair conclusions (Balagopalan et al., 2022). In high-risk situations, e.g., healthcare, the resulting non-interpretable errors could reduce a model's usefulness, making inherently interpretable models an appropriate choice. Nonetheless, there are domains where a fully interpretable model may not be necessary.

Building fully interpretable models is challenging. First, high-quality interpretable models may require extra time and effort from analysts with domain expertise compared to non-interpretable alternatives. Also, they sometimes fail to uncover "hidden patterns" within the data that black-box (i.e., non-interpretable) models may specialize in finding (Rudin, 2019). Doshi-Velez and Kim (2017) suggest interpretability is unnecessary when there is no significant impact or severe consequences for incorrect results or the problem is so well-studied and validated in real applications that one may trust the system's decisions.

We posit that defining interpretability as a binary notion is severely limiting. Instead, we define it as a domain-informed and user-defined parameter, allowing for models with varying levels of interpretability, capable of extracting the benefits of complex models but retaining interpretability for higher-risk predictions. However, this objective hinges on accurately anticipating such high-risk cases, e.g. those related to decisions concerning non-privileged groups in protected classes. This basis should allow for models that better balance interpretability, fairness, and performance trade-offs, avoiding focus on a single one.

Contributions. We introduce **FairMOE**, a Mixture of Experts (MOE) architecture using interpretable and non-interpretable experts, where a single expert is chosen per prediction. To the traditional MOE architecture we add (1) *Performance meta-learners* to anticipate the probability of a given expert prediction being correct; (2) a *Counterfactual Fairness Module* to identify highest-risk samples and ensure they are handled fairly, and; (3) an *Assignment Module* for expert selection, using results from the previous components within constraints of maximum levels of non-interpretability, i.e., the maximum amount of predictions from non-interpretable experts.

2 Related work

Our FairMOE proposal intersects four topics: (1) interpretability: definitions and contradictions; (2) mixture of experts, the basis for our proposal; (3) meta-learning, and how to anticipate predictive performance, and; (4) fairness and how to improve interpretability, fairness, and predictive performance trade-offs. Additionally, we evaluate how FairMOE handles high-risk samples and consider the various previous attempts at measuring sample risk.

Interpretability. XAI has rapidly increased in popularity. Arrieta et al. (2020) describe techniques involving transparent, inherently interpretable models and post-hoc explainability. Despite the popularity of post-hoc explainable models, there are many contexts when inherently interpretable models are superior. Rudin (2019) argues against explaining black box models for high-stakes decisions and demonstrates many of the flaws of explaining black box models. Carvalho et al. (2019) establish motivations for interpretability, including the potential impact of high-stakes decisions, societal concerns, and regulation. Despite a significant level of research, there is still no single agreed-upon definition of interpretability. Miller (2019) defines interpretability as “the degree to which a human can understand the cause of a decision”, while Kim et al. (2016) define it as “the degree to which a human can consistently predict the model’s result”.

One consistency is the idea that models are either interpretable or not. Recently, Frost et al. proposed a hybrid approach to interpretability in which a simple interpretable model will either make a prediction or will pass and allow a black box model to predict instead (Frost et al., 2024). In this paper, we use a continuous notion of interpretability, envisioning an architecture capable of minimizing the number of non-interpretable errors.

Mixture of Experts. Proposed over 30 years ago, MOE Jacobs et al. (1991) has been extensively explored within regression and classification tasks (Yuksel et al., 2012). Recently, sparse MOE has been used as layers to large neural networks (Shazeer et al., 2017) and as a vision transformer (Riquelme et al., 2021) to increase large, deep learning tasks’ efficiency. Closer to our work, Ismail et al. (2022) applied an interpretable MOE approach to structured and time series data, using an Assignment Module to pick individual expert for predictions and variable percentage of samples assigned to interpretable experts. Our approach leverages meta-learners to predict the accuracy of each expert given a specific sample, inspired by Cerqueira et al. (2017) work on time series forecasting.

Meta-Learning. Meta-learning has been applied to domains such as transfer learning, neural networks, and few-shot learning (Vanschoren, 2018). In each of them, it is used for error correction. The model’s weights are adjusted based on the knowledge gained from meta-features. We use meta-learning for error anticipation, towards selecting the best model. Khan et al. (2020) detail meta-learners’ usage for classifier selection. In an error-anticipation context, meta-learners are trained to predict model performance using a combination of the original feature space, meta-features, and model predictions. Using meta-learners, our proposal creates a fully-interpretable pipeline for selecting individual models and allows us to exploit each model’s strengths. However, by optimizing our proposal for predictive performance, this might create additional issues with regard to model fairness.

Fairness. There are two main approaches to analyzing fairness. Group fairness measures disparate treatment in protected groups over predictions, including pre-processing, in-processing, and post-processing methods (Hort et al., 2022). Pre-processing includes methods such as relabeling data (Kamiran & Calders, 2012), perturbation, and sampling (Chakraborty et al., 2021). Post-processing methods include input correction (Adler et al., 2018), classifier correction (Hardt et al., 2016), and output correction (Kamiran et al., 2012). Xian et al. (2023) explore the trade-off of fairness and performance and propose a post-processing algorithm using fair classifier score functions. In-processing methods attempt to train a model to learn fairness concepts. Agarwal et al. (2018) use adversarial learning. Zafar et al. (2017) apply constraints to the loss function to ensure fairness. Other approaches include a composition of multiple classification models (Pleiss et al., 2017) and adjusted learning (Zhang et al., 2021). Fairness can also be measured on an individual or sample-wise basis. Kusner et al. (2017) proposed the notion of counterfactual fairness, which uses the tools from causal inference to establish a prediction as fair if an individual’s prediction remains the same with changing protected attributes. Counterfactual fairness has been adopted in several domains as a viable approach toward fairness. For example, Garg et al. (2019) apply counterfactual fairness to text classification by considering perturbations obtained by substituting words within specific identity groups. Meanwhile, Guo et al.

applied counterfactual fairness to Graph Neural Networks to learn fair node representations for node classification tasks (Guo et al., 2023). Our approach uses counterfactual fairness to ensure our selected model predicts samples consistently. That is, selected experts should not discriminate against different protected attribute values. We separately evaluate our results with group fairness.

Assessing Risk of Examples Attempts at identifying difficult to predict cases have long been a focus in Imbalanced Learning literature. Frequently, cases are defined as either safe or unsafe where safe examples are those whose nearest neighbors make up entirely one class (Han et al., 2005; Kubat & Matwin, 1997; Laurikkala, 2001; Stefanowski, 2013). In some cases, unsafe examples are further classified as borderline, noisy, rare, or outlier (Kubat & Matwin, 1997; Napierala & Stefanowski, 2016) to further distinguish between levels of risk. Similarly, topics such as one-class novelty prediction attempt to identify test examples that do not belong to the train distribution (Ruff et al., 2018). Perera et al. (2019) use latent representations of in-class examples to detect unique classes. Ding et al. (2022) use KNN and Generative Adversarial Networks as part of a new sampling approach to detect intrusions across a network.

In our proposal, we perform post-hoc analysis using the safe, borderline, rare, and outlier categories proposed by Napierala and Stefanowski (2016). Our analysis gauges how effectively FairMOE handles high-risk cases and opens up avenues for future research to ensure these are predicted using interpretable experts.

3 Fair mixture of experts

This section describes our fairness-aware MOE-based proposal. FairMOE has four main components: (1) individual experts, where each predicts each sample; (2) performance meta-learners, which predict the probability of each expert's prediction accuracy; (3) a counterfactual fairness model, to assess predictive consistency regardless of protected attribute values in each case and, (4) an assignment module, combining the outcome of the previous two components and solving for non-interpretable model usage constraints. This high-level workflow is illustrated in Fig. 1, and components are described below.

3.1 FairMOE components

(1) Experts. FairMOE leverages a set of diverse expert learners trained using half the training data, including interpretable and non-interpretable models.

(2) Performance Meta-learner. A performance meta-learner per expert is trained to predict the probability of an accurate prediction. For interpretability, meta-learners use one of the following algorithms: Logistic Regression, Naive Bayes, Decision Tree, or K-Nearest Neighbors. They are trained using 10-fold cross-validation with grid search. The expert prediction is included as a feature within training, and the ground truth is a binary value indicating whether the expert correctly classified the sample. The learners are fit using the unused half of the training data to ensure they are trained using out-of-sample predictions.

(3) Counterfactual Fairness Module. To assess the fairness of individual models in a given sample, FairMOE uses a counterfactual fairness approach inspired by Kusner et al. (2017). Let A, X, Y and (U, V, F) represent protected features, remaining features, the output of interest, and a causal model where U is a set of latent background variables, V a

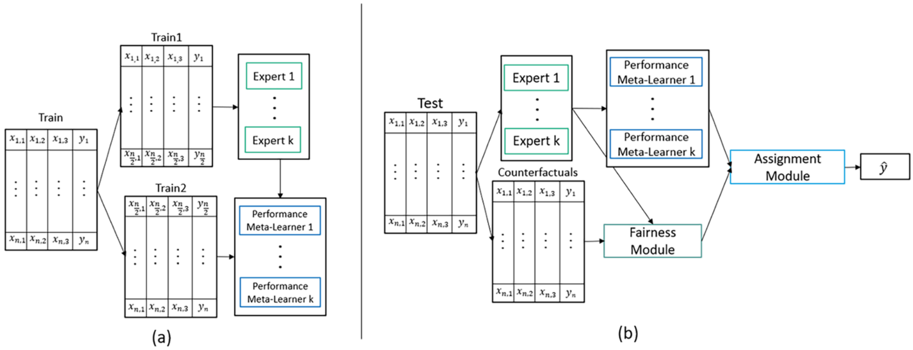

Fig. 1 **a** FairMOE training. Train data is split into two halves: *Train1* and *Train2*. Experts are trained with *Train1* and performance meta-learners on *Train2* using experts' predictions. **b** FairMOE testing. Experts predict the test data, which feeds into the respective performance meta-learners. Counterfactuals are generated around the protected attributes and assessed for consistency regarding expert predictions (*Fairness Module*). Finally, the *Assignment Module* uses the output from the *Fairness Module* and Performance Meta-Learners to select an expert and make the final prediction

set of observable variables, and F a set of structural equations. We define a predictor $\hat{Y}$ as counterfactually-fair if:

$$P(\hat{Y}_{\mathcal{A}\leftarrow\alpha}(U) = y|X = x) = P(\hat{Y}_{\mathcal{A}\leftarrow\alpha'}(U) = y|X = x) \tag{1}$$

where $\mathcal{A}$ is the set of all possible combinations of values within A, α is a given combination, and y is a given label.

The *Counterfactual Fairness Module* creates counterfactuals per sample to assess models with regard to individual (counterfactual) fairness. It generates a counterfactual for all possible permutations of Privileged/Unprivileged across protected classes, minus the original sample combination while holding all non-protected features constant.[1] Each expert then predicts them, being evaluated with a consistency score to determine their level of fairness, defined as:

$$CS = \frac{\sum_{\alpha'\in\mathcal{A}\setminus\alpha} I(\hat{Y}_{\mathcal{A}\leftarrow\alpha}(U)|X = x, \hat{Y}_{\mathcal{A}\leftarrow\alpha'}(U)|X = x)}{|\mathcal{A}| - 1}, \tag{2}$$

where I is an indicator function returning one if the two values match and 0 otherwise. Then, the Module selects the set of models per sample with the maximum consistency score:

$$M_x = \{\forall e \in E : CS(e) = max(CS(E))\} \tag{3}$$

where E is the set of experts. This set of counterfactually fair experts M_x is then used by the *Assignment Module* to pick the best fair expert per prediction.

(4) Assignment Module. Finally, the *Assignment Module* considers the performance meta-learners, the *Fairness Module*, and the *Non-Interpretable Budget* to select an expert

[1] Counterfactuals are created by (1) binarizing the features as privileged/unprivileged, (2) creating the permutations as described, and (3) transforming the binary value into a categorical or continuous variable by picking a random value following the distribution from the original training data.

for each sample. It has two stages. In the first stage, each sample within the test data is considered individually. The *Fairness Module* returns the fair experts for each sample to be predicted. Using the performance meta-learners, vectors are created with the interpretable and non-interpretable experts with the highest probability of an accurate prediction (HPAP), and the difference between the probabilities is calculated. In the second stage, the test data is considered as a whole. Samples with the highest positive difference, i.e., the probability of an accurate prediction is higher for the non-interpretable model, are assigned to the non-interpretable expert until the budget is exhausted. All remaining samples are assigned to the interpretable expert. Samples with a negative difference are always assigned to the interpretable expert, so the total budget is not always used. The designated expert's prediction is the final FairMOE prediction. The selection procedure is described in Algorithm 1.

Algorithm 1 Assignment Module

```
Require: $L$: Meta-Learners, $E$: Experts, $b$: Budget, $X$
for $x \in X$ do
    $M_x$ = FairnessModule($x$, $E$, $L$)
    $M^I \leftarrow \text{HPAP}(M_x \cap E_I)$, $E_I \in E$: subset of interpretable experts
    $M^{NI} \leftarrow \text{HPAP}(M_x \cap E_{NI})$, $E_{NI} \in E$: subset of non-interpretable experts
    $\Delta\text{prob} \leftarrow L_{M^{NI}}(x) - L_{M^I}(x)$
end for
$PNI$ = SelectPositive($\Delta$prob, b)
return $x \in X$: if $x \in PNI$: $M^{NI}(x)$; otherwise $M^I(x)$
```

4 Experimental evaluation

First, we present the data and methods used. Then, we proceed to assess the performance of FairMOE regarding predictive accuracy, interpretable decision-making, and fair behavior. We compare such performance against state-of-the-art baselines, aiming to answer the following research questions:

RQ1 Does the *Non-Interpretable Budget* impact predictive performance?
RQ2 Does FairMOE improve the predictive performance and fairness trade-off?
RQ3 What is the impact of the *Counterfactual Fairness Module*?
RQ4 Does FairMOE scale well with larger data sets?
RQ5 How well does FairMOE$_{1.0}$ assign high-risk predictions to Interpretable experts?

4.1 Data

We use nine fairness-oriented and public data sets (Le Quy et al., 2022) (Table 1), following the pre-processing steps, protected class definitions, and privileged groups described in Le Quy et al. (2022). For all data sets, pre-processing included removing samples which

Table 1 Data sets used in the experimental evaluation

Name	Prediction task	Cases	Feat	Protected attributes	Privileged classes
Adult (Dua & Graff, 2017)	Annual income exceeds $50,000	45222	94	Sex, race, age	Male, white, 25–60
German credit (Dua & Graff, 2017)	Bank Account is high credit risk	1000	47	Sex, age	Male, 25+
Dutch census (Center, 2013)	Person's occupation is prestigious	60420	50	Sex	Male
Bank marketing (Moro et al., 2014)	Client subscribes with deposit	45211	42	Age, marital status	25–60, Married
Credit card clients (Yeh & Lien, 2009)	Client will default in next month	30000	82	Sex, marital status	Male, single
OULAD (Kuzilek et al., 2017)	Student will pass class	21562	40	Sex	Male
Lawschool (Wightman, 1998)	Student will pas bar on first attempt	20798	18	Sex, race	Male, white
Diabetes (Strack et al., 2014)	Patient readmits within 30 days	173	272	Sex	Male
KDD census (Census-Income (KDD), 2000)	Annual income exceeds $50,000	284556	443	Sex, race	Male, white

had missing data and dropping non-predictive columns such as IDs. When necessary, the target variable was converted to a binary value and categorical variables were one-hot encoded. We used the definitions from Le Quy et al. (2022) for the privileged groups. When undefined, the majority class was designated as the privileged group.

4.2 Algorithms

We compare FairMOE against each expert and six fairness-aware algorithms. To build FairMOE we used seven algorithms as experts, optimized using grid search with 10-fold cross-validation (Table 2): Logistic Regression, Decision Tree, Naive Bayes, K-Nearest Neighbors (KNN) are interpretable, and Random Forest, LightGBM (LGBM), and XGBoost (XGB) are not.

Concerning fairness-aware algorithms, we used the solutions proposed by Hardt et al. (2016) (post-processing optimization of equalized odds), Zafar et al. (2017) (builds models using covariance between a sample's sensitive attributes to measure the decision boundary fairness, which guarantees disparate impact's business necessity clause, by maximizing fairness subject to accuracy constraints), Agarwal et al. (2018) (reduces a fairness classification task to a series of cost-sensitive classification problems, where the final outcome is a randomized classifier optimized for the most accurate classifier subject to fairness constraints) and xFAIR (Peng et al., 2022) (aims to mitigate bias and identify its cause by relabeling protected attributes in test data through extrapolation models designed to predict protected attributes through other independent variables). Additionally, we consider the Random Forest Fair Trees method proposed by Pereira Barata et al. (2023) which defines a new fairness-based tree splitting criteria and Adversarial Debiasing (AdvDeb) proposed by Zhang et al. (2018) which uses adversarial learning to address fairness concerns.

Hardt et al., Agarwal et al., and AdvDeb algorithms are implemented using the Fairlearn python package (Bird et al., 2020). For xFAIR, we used a Decision Tree as the extrapolation model and a Random Forest as the classification model suggested in the original paper (Peng et al., 2022). The Zafar et al. baseline was implemented using a Logistic Regression loss function. Of these alternatives, only the method proposed by Zafar et al. is interpretable. We adapted the authors' code to allow for multiple protected classes as necessary. Protected classes were encoded as binary features for the baselines incompatible with categorical or continuous features. AdvDeb is only able to optimize for a single protected attribute at a time. To adjust for this, models were fit optimizing for each of the protected attributes and the model which performed best on the test set was chosen for each of the evaluation metrics giving it an advantage over the other baselines.

We evaluate six versions of our method. The most basic version (noted as "Mode") considers the experts as an ensemble that predicts the most common prediction from all experts. Next, we consider an ensemble method that prioritizes fairness over performance (noted as "FairMode") by using the *Counterfactual Fairness Module* and predicting the most common prediction from only the counterfactual-fairest models, i.e., with a maximum consistency score. Alternatively, we consider the Mixture of Experts approach using performance meta-learners without the *Counterfactual Fairness Module* to test the interpretability aspect of our proposal, noted as "MOE". Finally, our full proposal "FairMOE", combines performance meta-learners, the *Counterfactual Fairness Module* and the *Assignment Module*. For MOE and FairMOE, we examined non-interpretable budgets of 0% (fully interpretable model) and 100% (no interpretability constraints), noted as $\text{MOE}_{0.0}$, $\text{FairMOE}_{0.0}$, $\text{MOE}_{1.0}$, and $\text{FairMOE}_{1.0}$, respectively.

Table 2 Overview of the solutions used as benchmarks including their name, underlying model(s), parameters, and whether or not the solution is interpretable

Model	Underlying algorithm(s)	Tuning parameters	Interpretable?
Expert 1	Logistic regression	N/A	Yes
Expert 2	Decision tree	Max. depth: [3, 5, 10, 15], Min. Samples per leaf: [5, 10, 25]	Yes
Expert 3	Naïve Bayes	N/A	Yes
Expert 4	KNN	Weights: distance, Neighbors: [5, 9, 13, ..., 45]	Yes
Expert 5	Random forest	Estimators: [10, 50, 100, 250], Min. samples per leaf: [5, 10, 25]	No
Expert 6	LGBM	Estimators: [10, 50, 100, 250], Learning rate: [.001,.01,.1], Min. samples per leaf: [5, 10, 25]	No
Expert 7	XGB	Estimators: [10, 50, 100, 250], Learning rate: [.001,.01,.1], Max. dEPTH: [3, 5, 10]	No
Agarwal (Agarwal et al., 2018)	LGBM	Estimators: [10, 50, 100, 250], Learning Rate: [.001,.01,.1], Min. samples per leaf: [5, 10, 25]	No
Hardt (Hardt et al., 2016)	LGBM	Estimators: [10, 50, 100, 250], Learning rate: [.001,.01,.1], Min. samples per leaf: [5, 10, 25]	No
Zafar (Zafar et al., 2017)	Logistic regression	N/A	Yes
xFAIR (Peng et al., 2022)	Decision tree, random forest	N/A	No
Fair Trees (Pereira Barata et al., 2023)	Random Forest	N/A	No
AdvDeb (Zhang et al., 2018)	Neural Network	N/A	No
Mode	Experts 1–7	N/A	No
Fair mode	Experts 1–7	N/A	No
$\text{MOE}_{0.0}$	Experts 1–7	N/A	Yes
$\text{MOE}_{1.0}$	Experts 1–7	N/A	Partially
$\text{FairMOE}_{0.0}$	Experts 1–7	N/A	Yes
$\text{FairMOE}_{1.0}$	Experts 1–7	N/A	Partially

4.3 Evaluation metrics

For thoroughness, we evaluate our results with Accuracy, F1-score, and G-mean. To measure fairness, we used Statistical Parity (SP) (Cynthia et al., 2012) and Equalized Odds (EO) (Hardt et al., 2016). SP is derived from the legal doctrine of Disparate Impact (Davis, 2004) but disregards ground truth labels, while EO considers them (Le Quy et al., 2022).

FairMOE is evaluated by running each data set 10 times with different 80%/20% train-test splits. For each iteration, the models were ranked by performance across all five metrics. With multiple protected classes, EO and SP are calculated for each protected class.

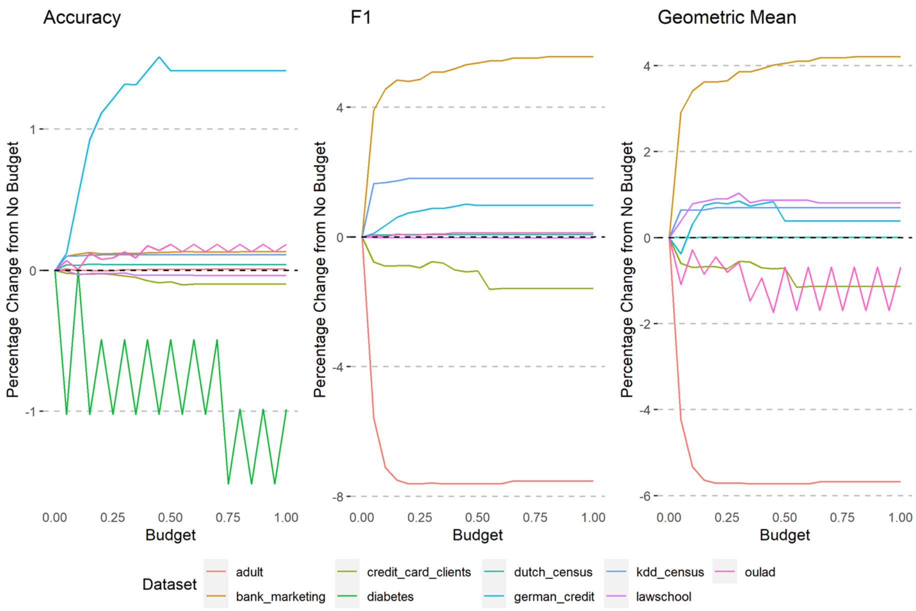

Fig. 2 Predictive performance of FairMOE at varying budgets. The performance lines represent average percentage change in Accuracy, F1, and G-Mean scores over ten runs compared to the fully interpretable FairMOE. Higher scores represent better performance. Note that the y-axes are not on the same scale

The metrics are grouped by performance (Accuracy, F1, G-mean) and fairness (SP, EO), and assessed as to the model's average ranking across these groups.

4.4 Results

4.4.1 Levels of interpretability (RQ1)

To measure the impact of interpretability on predictive performance, we test how Accuracy, G-Mean and F1 scores change as the *Non-Interpretable Budget* is increased (0–100% in 5pp) within each data set.

Results (Fig. 2) show that increasing the *Non-Interpretable Budget* can lead to predictive performance increases, but the magnitude of the effect is usually small. In all except one data set, the increase in accuracy is less than 1%. Additionally, in some cases increasing the use of more complex (non-interpretable) models worsens performance.

Importantly, results show that FairMOE performs well even in contexts where strict transparency is necessary. And, even when allowed to use the *Non-Interpretable Budget*, every metric quickly stabilizes when increasing the budget. We illustrate this in Fig. 3, showing that FairMOE does not need to resort to the total allotted non-interpretable predictions: with no interpretability constraints, FairMOE only used an average of 37.2% of the budget. This suggests that, in the majority of instances, fully interpretable models are capable of producing accurate predictions with high confidence. While non-interpretable models offer some performance benefits, these improvements occur on the margins

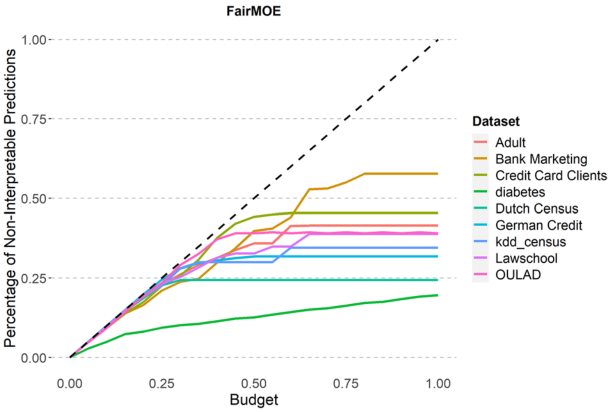

Fig. 3 Average total percentage of non-interpretable predictions for each budget in FairMOE and MOE. The dashed line indicates maximum budget usage

supporting our theory that peak performing models can be achieved while maintaining high interpretability.

4.4.2 Performance and fairness (RQ2)

Next, we compare how well FairMOE balances the predictive performance and fairness trade-off compared to other baselines, studying each baseline's Accuracy, F1-score, G-mean, SP, and EO rankings. The results depicted in Table 3 (grouped by metric type) show that:

1. Adding the *Counterfactual Fairness Module* notably increases group fairness at the cost of predictive performance;
2. Performance meta-learners add interpretability and fairness to our model with only a minor impact on predictive performance;
3. FairMOE is competitive with state-of-the-art baselines in predictive performance and fairness while increasing consistency and adding interpretability;
4. The *Non-Interpretable Budget* increases FairMOE's predictive performance without sacrificing fairness, demonstrating a cumulative advantage.

On predictive performance, XGB and LGBM are the best individual experts. While both are competitive with FairMOE overall, they produce non-interpretable models and poorly balance fairness and predictive performance (see the rightmost column in Table 3), limiting their utility in domains with fairness concerns. As for fairness-aware approaches, Fair Trees is the top model in group fairness. However, it is the weakest model in predictive performance.

Agarwal is competitive with FairMOE in both predictive performance and fairness and is the only model that outperforms FairMOE when considering the trade-off. Regardless, Agarwal's performance is less consistent than FairMOE with significantly higher standard deviations, and, importantly, Agarwal produces a non-interpretable model.

FairMOE, with and without interpretability constraints, shows competitive performance with regard to predictive and fairness. Figure 4 shows the magnitude of between model disparity with regard to predictive power beyond their rankings. FairMOE and MOE are

Table 3 Average and standard deviation of rankings (R) by predictive performance and fairness metrics

Predictive performance			Group fairness			All		
Solution	$\overline{R}$	$sd(R)$	Solution	$\overline{R}$	$sd(R)$	Solution	$\overline{R}$	ΔR
Agnostic								
Logistic regression	8.17	4.00	Logistic regression	12.28	4.30	Logistic regression	10.22	4.11
Decision tree	9.72	4.53	Decision tree	10.41	4.71	Decision tree	10.06	0.69
Naive Bayes	10.93	6.29	Naive Bayes	14.84	4.80	Naive Bayes	12.89	3.91
KNN	14.27	3.39	KNN	10.21	5.23	KNN	12.24	– 4.07
Random forest	9.53	4.63	Random forest	9.29	4.02	Random forest	9.41	– 0.23
LGBM	5.63	4.39	LGBM	11.73	4.33	LGBM	8.68	6.09
XGB	**5.05**	**3.74**	XGB	13.40	3.66	XGB	9.22	8.35
Aware								
Agarwal (Agarwal et al., 2018)	8.77	4.89	Agarwal	7.23	5.23	**Agarwal**	**8.00**	**– 1.54**
Hardt (Hardt et al., 2016)	12.00	7.72	Hardt	8.87	6.46	Hardt	10.43	– 3.13
Zafar (Zafar et al., 2017)	11.56	6.38	Zafar	10.11	6.83	Zafar	10.83	– 1.45
xFAIR (Peng et al., 2022)	10.30	4.98	xFAIR	9.12	4.38	xFAIR	9.71	– 1.18
Fair trees (Pereira Barata et al., 2023)	15.19	4.45	**Fair Trees**	**3.77**	**3.80**	Fair trees	9.48	– 11.42
AdvDeb* (Zhang et al., 2018)	14.05	5.63	AdvDeb*	7.41	6.63	AdvDeb*	10.73	– 6.64
Proposal								
Mode	7.66	3.77	Mode	11.27	4.17	Mode	9.46	3.61
Fair mode	10.59	3.93	Fair Mode	6.43	3.92	Fair mode	8.51	– 4.17
$\mathrm{MOE}_{0.0}$	8.94	3.79	$\mathrm{MOE}_{0.0}$	10.24	4.17	$\mathrm{MOE}_{0.0}$	9.59	1.2
$\mathrm{MOE}_{1.0}$	8.24	4.00	$\mathrm{MOE}_{0.0}$	10.63	4.03	$\mathrm{MOE}_{1.0}$	9.44	2.39
$\mathrm{FairMOE}_{0.0}$	9.90	3.38	$\mathrm{FairMOE}_{0.0}$	7.80	3.55	$\mathrm{FairMOE}_{0.0}$	8.85	– 2.10
$\mathrm{FairMOE}_{1.0}$	9.49	3.56	$\mathrm{FairMOE}_{1.0}$	7.25	3.44	$\mathrm{FairMOE}_{1.0}$	8.37	– 2.25

Bold values indicate the best performing algorithm by average rank

“All" is the mean of all predictive and fairness metric rankings and ΔR their difference. Solutions are grouped by Fairness Agnostic, Fairness Aware, and our proposal. Lower rankings signal better performance

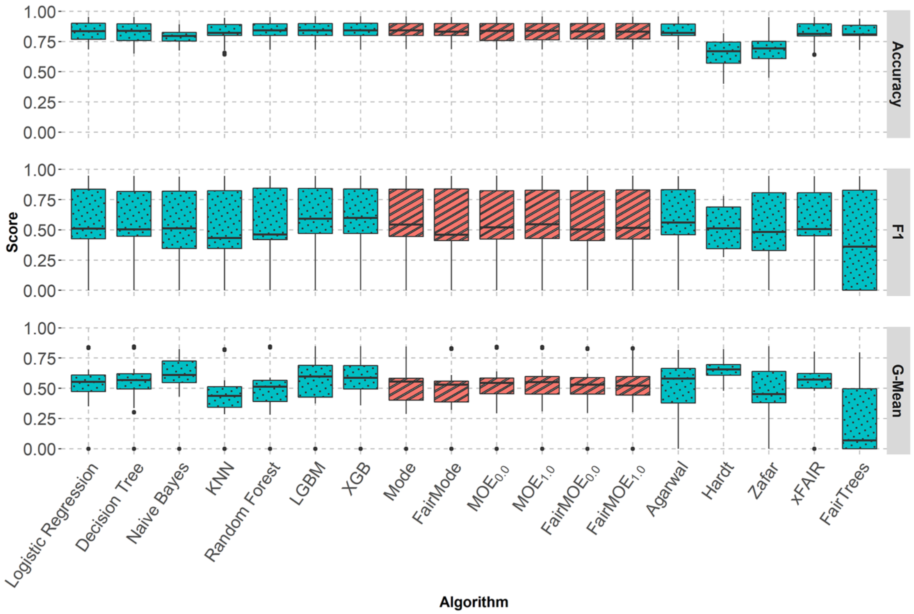

Fig. 4 Accuracy, F1, and G-mean scores per solution across all trials. Higher scores signal better performance

consistently in the middle or top-half of the accuracy and F1 box plots, suggesting they are competitive with the baselines. This is also observed concerning fairness metrics (Fig. 5).

Ultimately, FairMOE is competitive with the state-of-the-art baselines at striking a balance between fairness and predictive performance and can do so while maintaining interpretability. Even for high-risk domains, results show that a fully interpretable FairMOE (FairMOE$_{0.0}$) is competitive with baselines.

4.4.3 Counterfactual fairness module (RQ3)

Comparing the results of *Mode* and FairMode (Table 3), it is evident that the *Counterfactual Fairness Module* improves group fairness. *Mode* is one the worst-performing fairness-aware model regarding group fairness and is only fairer than the most extreme fairness-agnostic methods. Meanwhile, FairMode is the second-best trailing only Fair Trees. On the other hand, FairMode is weak in terms of predictive performance while Mode is the third-best, demonstrating the significant trade-off between fairness and predictive performance. The differences between MOE and FairMOE further support these findings. However, adding the performance meta-learners mitigates the loss in predictive performance. Additionally, removing the interpretability constraints from MOE, leads to a significant drop in group fairness as predictive performance is prioritized. However, in FairMOE, the model is able to maintain roughly equivalent levels of fairness and predictive performance. Overall, the *Counterfactual Fairness Module* successfully improves fairness while the performance meta learners add predictive performance and interpretability.

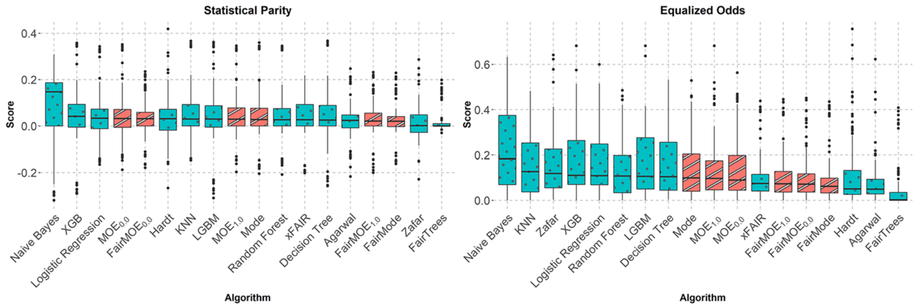

Fig. 5 SP and EO for each solution across all trials. Lower scores represent better performance. Note that the y-axes are not on the same scale

4.4.4 Scalability (RQ4)

Results show that FairMOE is competitive in predictive performance and fairness with regard to state-of-the-art baselines while producing consistent results. However, scalability is key. Table 4 shows the average total train and predict time per fairness-aware model, by data set. It shows that, while FairMOE is slower than Hardt and xFAIR, it improves over both alternatives in combined predictive performance and group fairness. Also, FairMOE is faster than Agarwal, the leading fairness-aware algorithm in most cases. Finally, Zafar, the only interpretable fairness-aware baseline, is much slower than other benchmarks and does not scale well. Overall, FairMOE is competitive with state-of-the-art baselines in terms of fairness and predictive performance trade-off, interpretable, faster, and more scalable than some of the leading alternatives.

4.4.5 Expert assignment (RQ5)

Following the methodology proposed by Napierala and Stefanowski (2016), we use a 5-Nearest Neighbor algorithm to classify each sample in the testing set as either Safe, Borderline, Rare, or Outlier based on the number of neighbors belonging to the same class label. We examine how frequently each group is being assigned to an interpretable or non-interpretable expert and their relative accuracies. These results are in Table 5. The Diabetes data set is omitted from this analysis because FairMOE consistently picked all interpretable experts.

Ideally, we would expect to see the majority of high-risk cases (rare and outlier) being handled by interpretable experts where the predictions can be easily audited. Meanwhile, safe and borderline cases would preferably be predicted by the most accurate expert. As demonstrated in RQ1, in all data sets except Bank Marketing, the interpretable experts are selected for a majority of the predictions. Notably, the Bank Marketing data set also has the largest drop in performance between the interpretable and non-interpretable experts.

Across all 8 data sets, the majority of outlier and rare samples are being assigned to interpretable experts suggesting that FairMOE is effective at ensuring high-risk predictions are interpretable. Additionally, while the non-interpretable models consistently have higher total accuracy, this is largely due to them being assigned a larger proportion of safe samples than the interpretable models. Despite non-interpretable models perceived ability to identify hidden patterns within the data, in 6 out of 8 data sets the interpretable models'

Table 4 Median processing time (seconds) to train and predict each solution in 10 trials per dataset

Dataset	Samples	Features	Median time (s)									
			Agarwal	Hardt	*Zafar	xFAIR	FairTrees	AdvDeb	*MOE$_{0.0}$	MOE$_{1.0}$	*FairMOE$_{0.0}$	FairMOE$_{1.0}$
German credit	1000	47	5.76	0.50	12.16	1.11	1.93	0.05	10.85	10.84	11.46	11.47
Lawschool	20798	18	50.47	1.14	102.47	10.18	17.62	0.22	25.58	25.44	38.16	38.25
OULAD	21562	40	32.7	1.08	114.11	5.24	7.02	0.25	25.89	25.95	28.18	28.03
Credit card clients	30000	82	45.80	4.03	782.62	23.57	38.09	0.45	45.65	45.69	68.67	68.99
Bank marketing	45211	42	121.55	3.19	648.35	27.32	28.56	0.63	56.82	56.88	90.31	90.54
Adult	45222	94	325.67	3.43	1648.83	41.81	27.45	0.80	67.79	67.79	121.37	121.76
Dutch census	60420	50	86.35	2.74	1051.52	15.66	16.29	3.55	73.98	73.95	79.74	79.45
Diabetes	173	272	1.94	0.37	71.47	0.58	1.15	0.06	12.75	12.75	12.86	12.86
KDD census	284556	443	1054.82	13.92	503716.00	1309.51	220.33	10.84	3431.46	3429.81	4345.09	4336.44

Asterisks (*) denote fully interpretable solutions

Table 5 Average assignment frequency and accuracy of experts separated by riskiness of sample. Higher values are italicized.

		Percentage assigned					Accuracy				
		Total	Safe	Borderline	Rare	Outlier	**Total**	Safe	Borderline	Rare	Outlier
Dutch Census	Int.	.756 ± .060	*.545 ± .144*	*.138 ± .068*	*.042 ± .020*	*.031 ± .011*	.844 ± .015	*.977 ± .004*	.676 ± .015	.249 ± .024	.073 ± .028
	NI	.244 ± .060	.144 ± .054	.068 ± .015	.020 ± .004	.011 ± .003	.781 ± .054	.958 ± .028	*.683 ± .029*	*.262 ± .074*	*.125 ± .065*
Adult	Int.	.518 ± .085	.314 ± .078	*.145 ± .007*	*.040 ± .004*	*.019 ± .002*	.790 ± .025	.961 ± .009	*.664 ± .010*	*.269 ± .015*	*.120 ± .029*
	NI	.482 ± .085	*.409 ± .078*	.049 ± .008	.014 ± .002	.010 ± .002	.912 ± .011	*.991 ± .002*	.642 ± .027	.240 ± .055	.050 ± .023
German Credit	Int.	.632 ± .109	*.272 ± .067*	*.256 ± .041*	*.080 ± .019*	*.025 ± .008*	.740 ± .042	.895 ± .038	*.718 ± .060*	*.401 ± .109*	*.412 ± .191*
	NI	.368 ± .109	.144 ± .055	.155 ± .056	.050 ± .016	.020 ± .008	.715 ± .068	*.910 ± .069*	.699 ± .098	.336 ± .144	.350 ± .178
Credit Card Clients	Int.	.565 ± .030	*.300 ± .023*	*.179 ± .009*	*.054 ± .003*	*.031 ± .002*	.791 ± .010	.957 ± .009	.746 ± .018	.339 ± .041	.236 ± .069
	NI	.435 ± .030	.287 ± .025	.094 ± .006	.031 ± .003	.023 ± .003	.859 ± .012	*.967 ± .008*	*.797 ± .021*	*.460 ± .055*	*.294 ± .068*
Bank Marketing	Int.	.376 ± .121	.254 ± .102	*.071 ± .017*	*.026 ± .004*	*.025 ± .003*	.797 ± .048	.969 ± .011	.674 ± .025	.274 ± .038	.108 ± .032
	NI	.624 ± .121	*.560 ± .102*	.043 ± .016	.011 ± .004	.010 ± .003	.960 ± .011	*.993 ± .002*	*.810 ± .041*	*.461 ± .053*	*.295 ± .094*
OULAD	Int.	.639 ± .034	*.240 ± .015*	*.292 ± .015*	*.078 ± .007*	*.029 ± .003*	.682 ± .008	.934 ± .011	.649 ± .014	*.243 ± .022*	*.104 ± .042*
	NI	.361 ± .034	.154 ± .016	.147 ± .013	.041 ± .005	.019 ± .003	.704 ± .014	*.955 ± .012*	*.666 ± .022*	.190 ± .053	.079 ± .035
Lawschool	Int.	.564 ± .173	*.458 ± .157*	*.068 ± .011*	*.021 ± .004*	*.018 ± .004*	.894 ± .017	.993 ± .002	*.718 ± .031*	.150 ± .043	*.028 ± .016*
	NI	.436 ± .173	.358 ± .158	.048 ± .011	.015 ± .003	.014 ± .004	.893 ± .024	*.994 ± .004*	.698 ± .053	*.159 ± .046*	.013 ± .019
KDD Census	Int.	.672 ± .089	*.611 ± .089*	*.039 ± .001*	*.012 ± .001*	*.010 ± .001*	.950 ± .008	.994 ± .001	.711 ± .015	.284 ± .018	*.082 ± .012*
	NI	.328 ± .089	.304 ± .089	.015 ± .001	.004 ± .000	.004 ± .000	.964 ± .008	*.997 ± .001*	*.778 ± .017*	*.295 ± .024*	.076 ± .015

Italic values indicates the higher values accuracy is within one standard deviation of the non-interpretables in predicting rare cases and 7 out of 8 in outlier cases. Overall, the performance of non-interpretable experts is similar to that of interpretable experts further supporting our finding in RQ1 that peak performance can be achieved without sacrificing interpretability when it is important. Future work will expand upon our current meta-learning technique. We envision finer-tuning of our method to create more hyper-specialized experts that minimize the number of high-risk non-interpretable predictions.

5 Discussion

This work intersects three essential concepts: predictive performance, fairness, and interpretability. The interactions between each of these are complex, and each has its own set of unique challenges.

FairMOE utilizes a *Non-Interpretable Budget* to address the trade-off between predictive performance and fairness. FairMOE balances the predictive performance of complex, non-interpretable models with the user-specified interpretability requirements with this budget. As our results demonstrate, FairMOE is capable of maintaining interpretability on more than 60% of predictions (average) without noticeable drops in performance. Further, our analysis demonstrates that FairMOE is effective in ensuring that high-risk cases are most commonly handled by interpretable experts. This finding illustrates that FairMOE is effective even in domains where strict interpretability is not necessary because it allows greater insight into the most-important, risky predictions without sacrificing predictive performance. Importantly, even in cases where strict interpretability is necessary, FairMOE performs competitively. This finding shows that FairMOE is applicable even in highly-regulated domains with strict transparency requirements. Introducing a user-defined, domain-specific *Non-Interpretable Budget* allows FairMOE to be amendable to different domain requirements.

Next, the *Counterfactual Fairness Module* within FairMOE addresses the trade-off between interpretability and fairness. By limiting our results to our counterfactually fair learners, FairMOE confines itself to making fair predictions even if such a result leads to a non-interpretable prediction. The results illustrate that, by adding the *Counterfactual Fairness Module*, we improve group fairness results.

Finally, we established FairMOE's success at balancing the predictive performance and fairness trade-off: it is the second-best option to Agarwal. To extend our understanding of how FairMOE handles this trade-off, in Fig. 6, we show how each solution performs with varying weights on performance and fairness. Our results show that FairMOE attains its success via consistent performance in both prediction and fairness whereas many of the baselines specialize in either one or the other. We make all the data and code available for reproducibility purposes at https://github.com/joegermino/FairMOE.

Limitations. Results are dependent on the number and diversity of the algorithms used for experts in training. In real-world applications, the definition of an acceptable level of interpretability is an open question that a domain expert will need to define based on their risk tolerance. Finally, evaluation was conducted measuring fairness on individual sensitive attributes. Ideal fairness measures should consider the effect of multiple sensitive attributes simultaneously. We envision future work exploring these topics and expanding our proposal to regression.

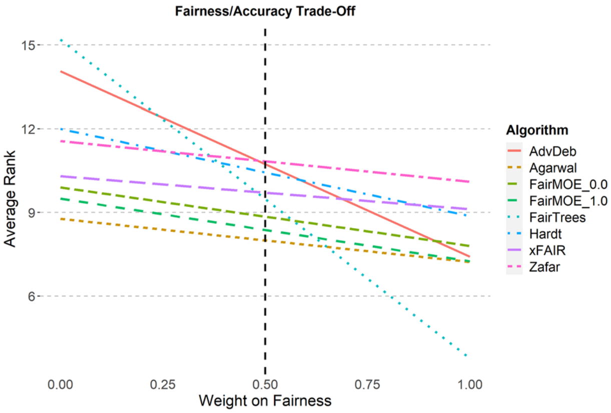

Fig. 6 The average global ranks of each fairness-aware baseline based on the weight given to fairness. Lower average ranks signal better performance

6 Conclusion

In this paper, we propose FairMOE, a fairness-aware solution based on the mixture of experts' architecture. Our proposal is the first to consider interpretability as a continuous, domain-informed notion. By combining three components: predictive meta-learners, the counterfactual fairness module, and the assignment module, we introduce a method which is able to achieve peak performance in the trade-off between predictive performance and fairness while maintaining high levels of interpretability. Our results demonstrate that the inclusion of a Non-Interpretable Budget allows for customizable levels of interpretability while improving overall performance. The counterfactual fairness module effectively improves group fairness performance without a significant reduction in predictive performance. Finally, we demonstrate that FairMOE is effective in identifying higher-risk cases that ideally would be handled by interpretable experts. Importantly, FairMOE challenges the paradigm that interpretability is a binary aspect of modeling. Instead, with FairMOE, we introduce the idea of interpretability as a continuous domain-informed notion that best exploits the typical performance interpretability trade-off.

Acknowledgements This research was supported by the Lucy Family Institute for Data and Society.

Author Contributions JG developed and designed the method and experiments. NM and NC provided significant mentoring and feedback throughout the project. JG managed the writing of this manuscript. All authors reviewed and approved the manuscript.

Data availability We make all the data and code available for reproducibility purposes at https://github.com/joegermino/FairMOE

Declarations

Conflict of interest The authors declare no conflict of interest.

References

Adler, P., Falk, C., Friedler, S. A., Nix, T., Rybeck, G., Scheidegger, C., Smith, B., & Venkatasubramanian, S. (2018). Auditing black-box models for indirect influence. *Knowledge and Information Systems, 54*(1), 95–122.

Agarwal, A., Beygelzimer, A., Dud ík, M., Langford, J., & Wallach, H. (2018). A reductions approach to fair classification. In *International conference on machine learning* (pp. 60–69). PMLR.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., & Benjamins, R. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion, 58*, 82–115.

Balagopalan, A., Zhang, H., Hamidieh, K., Hartvigsen, T., Rudzicz, F., & Ghassemi, M. (2022). The road to explainability is paved with bias: Measuring the fairness of explanations. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency* (pp. 1194–1206).

Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., & Walker, K. (2020). Fairlearn: A toolkit for assessing and improving fairness in AI. Microsoft, Tech. Rep. MSR-TR-2020-32.

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics, 8*(8), 832.

Census-Income (KDD). (2000). *UCI machine learning repository*. https://doi.org/10.24432/C5N30T .

Center, M. P. (2013). *Integrated public use microdata series international*. University of Minnesota Minneapolis.

Cerqueira, V., Torgo, L., Pinto, F., & Soares, C. (2017). Arbitrated ensemble for time series forecasting. In M. Ceci, J. Hollmén, L. Todorovski, C. Vens, & S. Džeroski (Eds.), *Machine learning and knowledge discovery in databases* (pp. 478–494). Cham: Springer.

Chakraborty, J., Majumder, S., & Menzies, T. (2021). Bias in machine learning software: Why? How? What to do? In *Proceedings of the 29th ACM joint meeting on European software engineering conference and symposium on the foundations of software engineering* (pp. 429–440).

Chandrasekaran, B., Tanner, M. C., & Josephson, J. R. (1989). Explaining control strategies in problem solving. *IEEE Intelligent Systems, 4*(01), 9–15.

Cynthia, D., Moritz, H., Toniann, P., Omer, R., & Richard, Z. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference* (pp. 214–226). Association for Computing Machinery, New York, NY, USA.

Davis, K. R. (2004). Age discrimination and disparate impact-a new look at an age-old problem. *Brook. L. Rev., 70*, 361.

Ding, H., Chen, L., Dong, L., Fu, Z., & Cui, X. (2022). Imbalanced data classification: A knn and generative adversarial networks-based hybrid approach for intrusion detection. *Future Generation Computer Systems, 131*, 240–254.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

Dua, D., & Graff, C. (2017). *UCI machine learning repository*. http://archive.ics.uci.edu/ml.

Frost, N., Lipton, Z., Mansour, Y., & Moshkovitz, M. (2024). Partially interpretable models with guarantees on coverage and accuracy. In *International conference on algorithmic learning theory* (pp. 590–613). PMLR.

Garg, S., Perot, V., Limtiaco, N., Taly, A., Chi, E. H., & Beutel, A. (2019). Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society* (pp. 219–226).

Guo, Z., Li, J., Xiao, T., Ma, Y., & Wang, S. (2023). Towards fair graph neural networks via graph counterfactual. In *Proceedings of the 32nd ACM international conference on information and knowledge management* (pp. 669–678).

Han, H., Wang, W.-Y., & Mao, B.-H. (2005). Borderline-smote: A new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing* (pp. 878–887). Springer.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, *29*, 3315–3323.

Hort, M., Chen, Z., Zhang, J. M., Sarro, F., & Harman, M. (2022). Bias mitigation for machine learning classifiers: A comprehensive survey. arXiv preprint arXiv:2207.07068.

Ismail, A. A., Arik, S.Ö., Yoon, J., Taly, A., Feizi, S., & Pfister, T. (2022). Interpretable mixture of experts for structured data. arXiv preprint arXiv:2206.02107.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation, 3*(1), 79–87.

Kamiran, F., Karim, A., & Zhang, X. (2012). Decision theory for discrimination-aware classification. In *2012 IEEE 12th international conference on data mining* (pp. 924–929). https://doi.org/10.1109/ICDM.2012.45.

Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems, 33*(1), 1–33.

Khan, I., Zhang, X., Rehman, M., & Ali, R. (2020). A literature survey and empirical study of meta-learning for classifier selection. *IEEE Access, 8*, 10262–10281.

Kim, B., Khanna, R., & Koyejo, O. O. (2016). Examples are not enough, learn to criticize! criticism for interpretability. *Advances in Neural Information Processing Systems,29*, 2280–2288.

Kubat, M., & Matwin, S. (1997). Addressing the curse of imbalanced training sets: One-sided selection. In: *Icml* (Vol. 97, p. 179). Citeseer.

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. *Advances in Neural Information Processing Systems*, *30*, 4066–4076.

Kuzilek, J., Hlosta, M., & Zdrahal, Z. (2017). Open university learning analytics dataset. *Scientific Data, 4*(1), 1–8.

Laurikkala, J.(2001). Improving identification of difficult small classes by balancing class distribution. In *Artificial intelligence in medicine: 8th conference on artificial intelligence in medicine in Europe, AIME 2001 Cascais, Portugal, July 1–4, 2001, Proceedings 8* (pp. 63–66). Springer.

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

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1–38.

Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems, 62*, 22–31.

Napierala, K., & Stefanowski, J. (2016). Types of minority class examples and their influence on learning classifiers from imbalanced data. *Journal of Intelligent Information Systems, 46*, 563–597. https://doi.org/10.1007/s10844-015-0368-1

Peng, K., Chakraborty, J., & Menzies, T. (2022). Fairmask: Better fairness via model-based rebalancing of protected attributes. *IEEE Transactions on Software Engineering*. https://doi.org/10.1109/TSE.2022.3220713

Pereira Barata, A., Takes, F. W., Herik, H. J., & Veenman, C. J. (2023). Fair tree classifier using strong demographic parity. *Machine Learning, 113*, 3305–3324.

Perera, P., Nallapati, R., & Xiang, B. (2019). Ocgan: One-class novelty detection using gans with constrained latent representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR).*

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. Q. (2017). On fairness and calibration. *Advances in Neural Information Processing Systems*, *30*, 5680–5689.

Riquelme, C., Puigcerver, J., Mustafa, B., Neumann, M., Jenatton, R., Susano Pinto, A., Keysers, D., & Houlsby, N. (2021). Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems, 34*, 8583–8595.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215.

Ruff, L., Vandermeulen, R., Goernitz, N., Deecke, L., Siddiqui, S. A., Binder, A., Müller, E., & Kloft, M. (2018). Deep one-class classification. In *International conference on machine learning* (pp. 4393–4402). PMLR

Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., & Dean, J. (2017). Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. arXiv preprint arXiv:1701.06538.

Stefanowski, J. (2013). Overlapping, rare examples and class decomposition in learning classifiers from imbalanced data. In *Emerging paradigms in machine learning* (pp. 277–306). Springer.

Strack, B., DeShazo, J. P., Gennings, C., Olmo, J. L., Ventura, S., Cios, K. J., Clore, J. N., et al. (2014). Impact of hba1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records. *BioMed Research International*, *2014*, 781670–781680.

Vanschoren, J. (2018). Meta-learning: A survey. arXiv preprint arXiv:1810.03548.

Wightman, L .F. (1998). Lsac national longitudinal bar passage study. lSAC research report series.

Xian, R., Yin, L., & Zhao, H. (2023). Fair and optimal classification via post-processing. In *International conference on machine learning*, (pp. 37977–38012). PMLR.

Yeh, I. .-C., & Lien, C. .-h. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications, 36*(2), 2473–2480.

Yuksel, S. E., Wilson, J. N., & Gader, P. D. (2012). Twenty years of mixture of experts. *IEEE Transactions on Neural Networks and Learning Systems, 23*(8), 1177–1193.

Zafar, M. B., Valera, I., Rogriguez, M. G., & Gummadi, K. P. (2017). Fairness constraints: Mechanisms for fair classification. In *Artificial intelligence and statistics* (pp. 962–970). PMLR.

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM conference on AI, ethics, and society* (pp. 335–340).

Zhang, W., Bifet, A., Zhang, X., Weiss, J. C., & Nejdl, W. (2021). Farf: A fair and adaptive random forests classifier. In *Pacific-Asia conference on knowledge discovery and data mining* (pp. 245–256).