Neural Networks Bias Mitigation Through Fuzzy Logic and Saliency Maps

Sahar Shah[a], Davide E. Ciucci[b], Sara L. Manzoni[c] and Italo F. Zoppis[d]
Department of Informatics, System and Communication, University of Milano-Bicocca, Viale Sarca 336, Milano, Italy
s.shah19@campus.unimib.it, {davide.ciucci, sara.manzoni, italo.zoppis}@unimib.it

Keywords: Bias Mitigation, Neural Networks, Decision-Making, Saliency Maps, Fuzzy Logic.

Abstract: Mitigating biases in neural networks is crucial to reduce or eliminate the predictive model's unfair responses, which may arise from unbalanced training, defective architectures, or even social prejudices embedded in the data. This study proposes a novel and fully differentiable framework for mitigating neural network bias using *Saliency Maps* and *Fuzzy Logic*. We focus our analysis on a simulation study for recommendation systems, where neural networks are crucial in classifying job applicants based on relevant and sensitive attributes. Leveraging the interpretability of a set of *Fuzzy implications* and the importance of features attributed by *Saliency Maps*, our approach penalizes models when they overly rely on biased predictions during training. In this way, we ensure that bias mitigation occurs within the gradient-based optimization process, allowing efficient model training and evaluation.

1 INTRODUCTION

Bias in machine learning remains a significant challenge, particularly in decision-making systems that involve *sensitive features* such as gender, race, or disability status. Recent research has shown that models trained on biased data can perpetuate or even exacerbate social prejudice, leading to unfair predictions (Kamiran et al., 2010). Various strategies have been proposed to address this problem, including pre-processing, in-processing, and post-processing methods.

Pre-processing methods involve removing bias from the dataset before it is used for training (Ghosh et al., 2023). Common techniques include adjusting the importance of data samples to ensure a balanced representation between sensitive groups (Kamiran and Calders, 2012) or editing feature distributions to reduce disparate impacts without compromising the utility of data (Feldman et al., 2015).

In-processing approaches focus on modifying the learning algorithm to incorporate fairness constraints during training (Iosifidis and Ntoutsi, 2019). For example, Kamiran et al. (2010) modify the splitting criterion of decision trees to consider the impact of the split on the protected attribute. Similarly, Kamishima et al. (2012) introduces a regularized technique to reduce the effect of indirect prejudice (measured as the mutual information between sensitive features and class labels). Furthermore, constraints, such as demographic parity or equalized odds, are directly integrated into the optimization objective in Zafar et al. (2017). Recent work on adversarial networks has also focused on minimizing predictive disparities between sensitive groups (Zhang et al., 2018).

Post-processing methods adjust predictions without altering the model or data. For example, Hardt et al. (2016) modify predictions to satisfy the fairness criteria, such as equalized odds, while (Pleiss et al., 2017) balances accuracy and fairness through calibrated prediction adjustments.

Although there has been extensive work on bias mitigation, the explicit integration of *eXplainable Artificial Intelligence* (XAI) techniques for bias reduction has been less explored particularly for in-processing methods (Tjoa and Guan, 2020). In this paper, we focus on such an integration aiming not only to enhance *explainability* (according to XAI) but also to allow the model to adjust dynamically predefined, human-understandable fairness constraints. This capability is critical, especially in high-risk applications such as medical diagnostics, where human intervention is essential.

[a] https://orcid.org/0009-0001-5588-8823
[b] https://orcid.org/0000-0002-8083-7809
[c] https://orcid.org/0000-0002-6406-536X
[d] https://orcid.org/0000-0001-7312-7123

Shah, S., Ciucci, D. E., Manzoni, S. L. and Zoppis, I. F.
Neural Networks Bias Mitigation Through Fuzzy Logic and Saliency Maps.
DOI: 10.5220/0013366800003890

In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence (ICAART 2025) - Volume 3*, pages 1343-1351
ISBN: 978-989-758-737-5; ISSN: 2184-433X

The intent is to modify the model's loss function by including a regularization term or constraints that account for discriminatory behavior or fairness criteria. By leveraging the importance of sensitive features, as attributed by *Saliency Maps*, we evaluate human-predefined rules (i.e., fuzzy antecedents of an implication). Simultaneously, we quantify the bias for neural loss regularization by evaluating the output of these rules (fuzzy consequent). In other words, fuzzy rules provide humanly understandable and readable expressions of how certain sensitive attributes (race, gender, and disability status) should or should not influence the prediction process. Importantly, this approach will ensure that bias mitigation occurs during gradient backpropagation, allowing efficient model training and evaluation.

The paper is organized as follows. Section 2 outlines the methods utilized in our analysis; Section 3 describes the numerical experiments conducted along with the insights obtained, and Section 4 concludes this paper with a discussion of potential future extensions.

2 MATERIALS AND METHODS

Although our approach is generalizable to other setting and applications, we focus on neural decision-making for classification of job applicants. A neural network categorizes applicants into two groups: those qualified for higher-skilled positions and those qualified for lower-skilled positions. The following sections outline the methods employed in our approach.

2.1 Saliency Maps

Saliency maps (SM) are part of a broader category of methods known as *attribute methods*, which provide insights into which attributes have the greatest influence on the corresponding predictive output. We focus on SM because they are intrinsically differentiable, meaning they exploit the gradients of the model's predictions $f(\mathbf{x})$ for each input (x), i.e.,

$$S(\mathbf{x}) = \left| \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} \right| \tag{1}$$

The differentiability of $f(\mathbf{x})$ allows us to apply a gradient-based optimization to the neural network custom loss function. Specifically, the gradients in Eq. 1 are evaluated and averaged across epochs to determine the importance of each sensitive feature in decision-making. To this end, the average sensitivity of the network responses (with respect to the sensitive features) is sliced from $S(\mathbf{x})$, and passed to the fuzzy controller, as described in the next paragraph, to guide the bias mitigation efforts.

2.2 Fuzzy Controller

A fuzzy logic system (referred to here as a *fuzzy controller*) refines our approach by incorporating human knowledge through a set of fuzzy rules. The controller's main goal is to quantify bias and ensure that relevant factors for decision-making are adequately addressed. This task is achieved by carefully tuning the linguistic terms (fuzzy sets) as specified by human operators. Furthermore, to effectively integrate the fuzzy logic and enable gradient backpropagation (in-processing), we have appropriately applied differentiable functions to approximate the typical fuzzy operators. We report the applied controller, using a typical pipeline for the fuzzy system design.

2.2.1 Fuzzification

Let $s \in [0,1]$ be the saliency value of any sensitive feature. To allow differentiation, we define the following Gaussian curves as membership functions for three fuzzy sets: Low, Moderate, and High.

$$\begin{aligned} \mu_{\text{Low}}(s) &= \exp\left(-\frac{(s-c_{\text{Low}})^2}{2\sigma^2}\right) \\ \mu_{\text{Moderate}}(s) &= \exp\left(-\frac{(s-c_{\text{Moderate}})^2}{2\sigma^2}\right) \\ \mu_{\text{High}}(s) &= \exp\left(-\frac{(s-c_{\text{High}})^2}{2\sigma^2}\right) \end{aligned} \tag{2}$$

Here $c_{\text{Low}} = 0, c_{\text{Moderate}} = 0.5, c_{\text{High}} = 1$ represent the Gaussian centers, while σ denotes the standard deviation.

2.2.2 Fuzzy Rules Definition

Given a pair of saliency values $s_1, s_2 \in [0,1]$, and an output variable y (representing bias), our rules follow a typical "if-then" implication with the format:

$$R_j : \text{IF } s_1 \text{ is } A_{j1} \text{ AND } s_2 \text{ is } A_{j2} \text{ THEN } y \text{ is } B_j \tag{3}$$

In Eq. 3 A_{j1} and A_{j2} represent a pair of fuzzy sets in rule R_j, y denotes the output variable (bias), and B_j is the output fuzzy set with Gaussian membership function (similarly to Eq. 2), respectively with center in $c_{\text{Out Low}}, c_{\text{Out Moderate}}, c_{\text{Out High}}$.

All rules applied in this study are reported in Tab. 1. To ensure a conservative and interpretable fuzzy inference process, the rules are consistent with a *fuzzy* *partial ordering*, i.e., the fuzzy sets representing linguistic terms (e.g., *Low*, *Moderate*, *High*) are ordered based on their membership dominance (Zadeh, 1971), establishing a natural hierarchy such that:

$$\text{Low} < \text{Moderate} < \text{High}. \tag{4}$$

Equation 4 reflects a progression in intensity (or magnitude), and ensures that the output adheres to the principle of being the *smallest fuzzy set* consistent with the defined order. This conservative approach prevents overestimation in the inference process by assigning the outputs that are minimal in terms of the defined hierarchy, preserving interpretability and caution.

2.2.3 Rule Evaluation and Aggregation

Each rule (Eq. 3) is evaluated by applying a Product-Over-Sum *softmin* function, defined as

$$\text{softmin}(\mu_{A_{j,1}}(s_1), \mu_{A_{j,2}}(s_2)) = \frac{\mu_{A_{j,1}}(s_1) \cdot \mu_{A_{j,2}}(s_2)}{\mu_{A_{j,1}}(s_1) + \mu_{A_{j,2}}(s_2)} \tag{5}$$

for any pair of fuzzy sets A_{j1}, A_{j2}. Eq. 5 approximates the minimum *firing strength*, α_j, for the rule R_j across the antecedents, thus providing smooth transitions between values. Importantly, since Eq. 5 depends on the sums and products of differentiable functions (Gaussian curves), it is itself differentiable, making it useful for gradient-based optimization.
Finally, we aggregate the bias across the rules using a Weighted Sum (softmax) of each Gaussian output, thus maintaining the differentiability of the fuzzy operations, i.e.,

$$\begin{aligned}\mu_{\text{agg}}(y) &= \text{softmax}(\mu_{B_1}(y), \mu_{B_2}(y), \ldots, \mu_{B_n}(y)) \\ &= \sum_{i=1}^{m} \alpha_i \cdot e^{-\frac{(y-c_i)^2}{2\sigma^2}}\end{aligned} \tag{6}$$

2.2.4 Defuzzification

The defuzzification converts the fuzzy output back into a quantitative measure which serves, in this case, as a quantitative measure of the detected bias. The well-known weighted average method (centroid) is considered, i.e.,

$$y^* = \frac{\int y \cdot \mu_{\text{agg}}(y)\, dy}{\int \mu_{\text{agg}}(y)\, dy} \tag{7}$$

When the Gaussian curves share the same deviation, then Eq. 7 simplifies to a weighted average of Gaussian centers (c_{Low}, c_{Moderate}, and c_{High}), over the sum of the activation weights,

$$y^* \approx \frac{\sum_i \alpha_i \cdot c_i}{\sum_i \alpha_i} \tag{8}$$

where c_i are the centers of the output Gaussian membership (i.e., Low, Moderate, High), and α_i are the *rule activations*.

To derive Eq. 8, for our application, it suffices to note that according to Eq. 6, we can expand the numerator of y^* as

$$\int y \cdot \mu_{\text{agg}}(y)\, dy = \int y \cdot \sum_{i=1}^{m} \alpha_i \cdot e^{-\frac{(y-c_i)^2}{2\sigma^2}}\, dy \tag{9}$$

Then, bringing the sum outside the integral,

$$\int y \cdot \mu_{\text{agg}}(y)\, dy = \sum_{i=1}^{m} \alpha_i \int y \cdot e^{-\frac{(y-c_i)^2}{2\sigma^2}}\, dy \tag{10}$$

Each integral $\int y \cdot e^{-\frac{(y-c_i)^2}{2\sigma^2}}\, dy$ represents the expected value of y for a Gaussian curve centered at c_i, or equivalently the expected value of a Gaussian distribution with parameters (c_i, σ), scaled by $\sqrt{2\pi\sigma^2}$, i.e.,

$$\int y \cdot e^{-\frac{(y-c_i)^2}{2\sigma^2}}\, dy = c_i\sqrt{2\pi\sigma^2} \tag{11}$$

Thus, the numerator becomes,

$$\sum_{i=1}^{m} \alpha_i \cdot c_i\sqrt{2\pi\sigma^2} \tag{12}$$

Similarly, by extending the denominator in Eq. 8, we get

$$\int \mu_{\text{agg}}(y)\, dy = \int \sum_{i=1}^{m} \alpha_i \cdot e^{-\frac{(y-c_i)^2}{2\sigma^2}}\, dy \tag{13}$$

bringing out the sum outside the integral,

$$\int \mu_{\text{agg}}(y)\, dy = \sum_{i=1}^{m} \alpha_i \int e^{-\frac{(y-c_i)^2}{2\sigma^2}}\, dy = \sum_{i=1}^{m} \alpha_i \cdot \sqrt{2\pi\sigma^2} \tag{14}$$

Finally, simplifying common terms in Eq. 12, and 14 we result with Eq. 8. The crisp value y^* is then incorporated into the overall loss to mitigate biased predictions.

2.3 Neural Network

The Neural Network is relatively simple: a feed-forward architecture that generates biased decisions only. Essentially, the model will be required to respond to sensitive and relevant input to determine

Table 1: Fuzzy rules used to determine bias levels based on saliency map values.

Fuzzy Rules	Descriptions
R_1	IF Saliency of Race is High AND Saliency of Gender is Medium THEN Bias is High
R_2	IF Saliency of Race is Medium AND Saliency of Gender is Low THEN Bias is Medium
R_3	IF Saliency of Race is Low AND Saliency of Gender is High THEN Bias is High
R_4	IF Saliency of Gender is High AND Saliency of Disability is Medium THEN Bias is High
R_5	IF Saliency of Gender is Medium AND Saliency of Disability is Low THEN Bias is Medium
R_6	IF Saliency of Gender is Low AND Saliency of Disability is High THEN Bias is High
R_7	IF Saliency of Disability is High AND Saliency of Race is Medium THEN Bias is High
R_8	IF Saliency of Disability is Medium AND Saliency of Race is Low THEN Bias is Medium
R_9	IF Saliency of Disability is Low AND Saliency of Race is High THEN Bias is High
R_{10}	IF Saliency of Race is Low AND Saliency of Gender is Low AND Saliency of Disability is Low THEN Bias is Low

which of the two jobs assigned classes a candidate belongs to (i.e., either higher-skilled or lower-skilled class). As mentioned previously, we follow in-processing approaches: the core mechanism of our predictions is a loss function, $\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda y^*$, that combined $\mathcal{L}_{\text{CE}}$, implemented as a typical binary cross-entropy, with the defuzzified bias y^* obtained in Eq. 8 .

This regularization should penalize during training the model when the antecedents of a fuzzy implication is met (i.e., when the saliency map shows the model overly relies on biased features, as returned by the fuzzy controller).

2.4 Data Set

We implemented a data generation mechanism that explicitly introduces bias through an unfair filter condition to discriminate against a specific demographic group. We define the following two sets of features in the simulated data.

- *Sensitive features* include Race, Gender, and Disability status. These are data attributes that require special protection due to their nature, and any bias introduced here could result in prejudice towards a group of applicants.
- *Relevant features* reflect the subject's Skills, Experience, and Education. These are legitimate factors for a fair job assignment and represent the quality levels in applicant profiles.

The generation process (reported in Tab. 1) correlates relevant features (mean score) with skillful job labels for every profile, and penalizes those profiles that match the discriminatory filter, reducing the job label assigned. The resulting label finally provides a job class.

Algorithm 1: Job Class Assignment.

```
Require:
  Profile $x$ (Sensitive & Relevant Features)
  Sensitive Filter:
  e.g., Race = 1, Gender = 0, Disability = 1
Ensure: Job class assignment (JClass)
  score $\leftarrow$ skill + exp + edu                    ▷ Fair correlation
  score $\leftarrow$ (score − score$_{\min}$)/(score$_{\max}$ − score$_{\min}$)
  jLab $\leftarrow \lfloor 6 \times \text{score} \rfloor + 1$          ▷ Mapping to integer
  if $x$ matches the filter then                     ▷ Penalization
    jLab = jLab − 2
    jLab = max(jLab, 0)                       ▷ ensure jLab $\geq 0$
  end if
  if $1 \leq$ jLab $\leq 3$ then
    jClass $\leftarrow$ 0             ▷ Lower-Skilled Job assigned
  else if $4 \leq$ jLab $\leq 6$ then
    jClass $\leftarrow$ 1            ▷ Higher-Skilled Job assigned
  end if
```

3 NUMERICAL EXPERIMENTS

We conducted a series of numerical experiments to evaluate the effectiveness of our approach, focusing mainly on the influence of sensitive features on neural decision making. To accomplish this goal, we designed our experiments to evaluate and estimate the mitigation gain through the introduction of a regularization term. In detail, the following three models are considered.

1. *Fuzzy-regularized model.* A model based on fuzzy regularization.
2. *Saliency-regularized model.* In this case, the regularization term is obtained by combining (averaging) the importance values assigned to feature across epochs (Saliency values).
3. *Non-regularized model - No regularization* is applied.

All models are simple, fully connected (dense) feed-forward networks designed for binary classification. They aim to establish a baseline for testing bias mitigation through regularization while ensuring that the models remain interpretable and efficient. All models share the following architectures and are trained with Adam Optimizer.

- *Input Layer*. Accepts a vector of six features as input: three relevant features (skill, experience, and education) and three sensitive features (race, gender, and disability status).
- *Hidden Layer*. A fully connected dense layer with 10 hidden units, with ReLU (Rectified Linear Unit).
- *Output Layer*. A fully connected dense layer with a single sigmoid output that is interpreted as the likelihood of the sample belonging to the higher-skilled job class (class 1). A threshold of 0.5 is applied to classify samples into class 0 or class 1.

To evaluate generalization and robustness, *10-fold cross-validation* is applied. Training meta-parameters such as batch size, percentages of training and testing, and learning rate are the same for all the experiments. Each model is trained for a fixed number of 200 epochs in each cross-validation fold. The choice of 200 epochs is made based on initial experiments, which suggested that the models achieve stable accuracy and loss convergence; no stop criteria are applied. For each cross-validation fold, we proceed as follows:

- The data is partitioned into 90% for training and 10% for validation.
- A new model is initialized and trained from scratch (for each fold).
- After each fold, the training and validation metrics are accumulated.

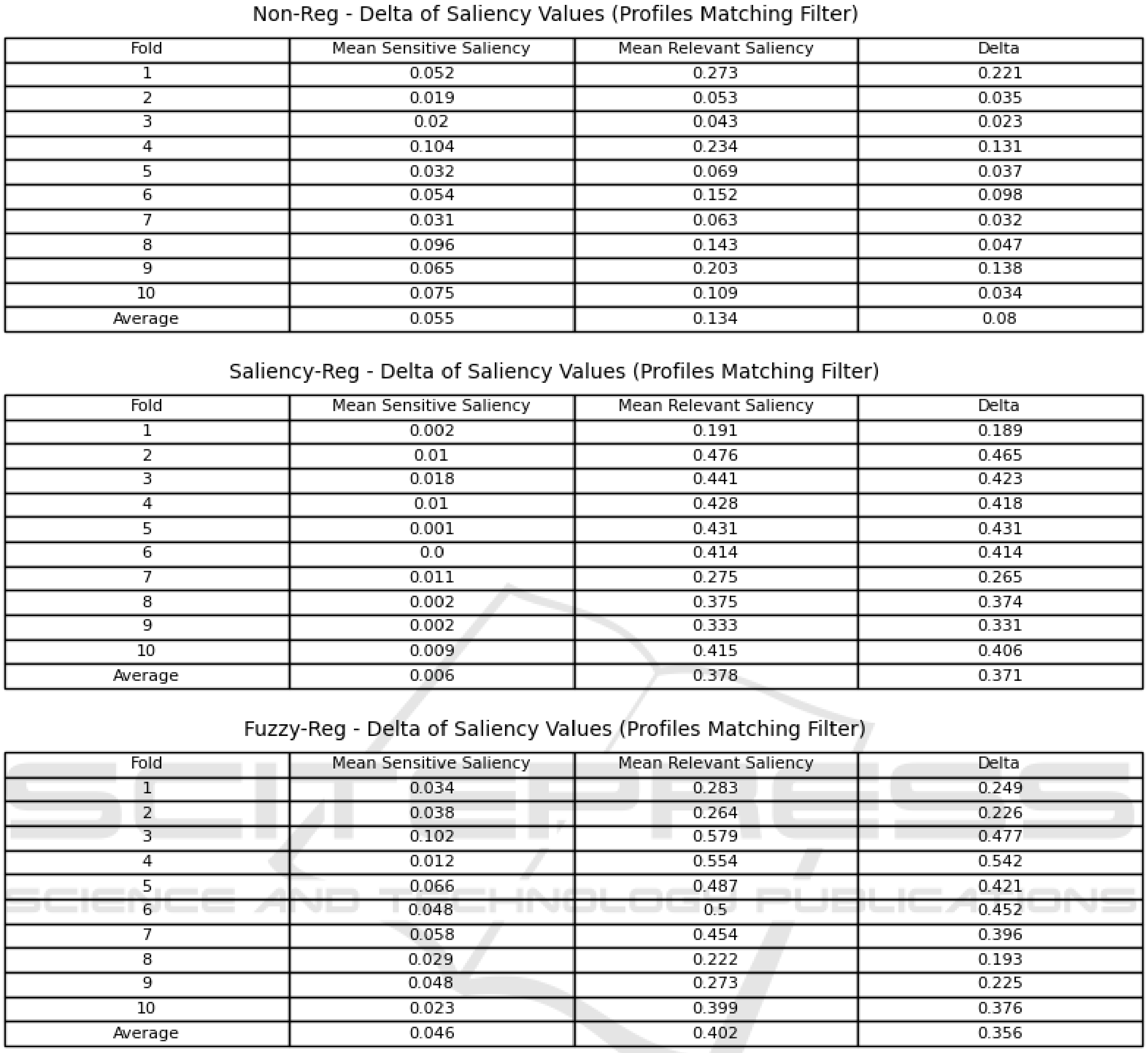
Moreover, Accuracy and Saliency values were collected for each sensitive and relevant feature for profiles that match the filter and those that do not.
To better interpret our results, it is important to note that we are focusing on classification tasks. The predictions made by our models will be compared to data obtained through an unfair generation process that assigns labels with inherent biases. As a result, comparing regularized and non-regularized models (under similar experimental conditions) may show a decrease in accuracy for the regularized model. This decrease may reflect the impact of bias mitigation efforts, which aim to promote fair decision-making that differs from the original biased labels assigned by the unfair data generation mechanism. In our experiments, we directly assessed the impact of regularization on saliency maps. The following paragraphs provide a detailed report of our analysis and the corresponding results.

3.1 Saliency Analysis

We directly assessed the impact of regularization on saliency maps. In particular, we evaluated the effect of model type on bias mitigation by estimating the amount of bias reduction as a dependent variable of ANOVA models.

To proceed, we initially considered the Null hypothesis that there is no difference in mean saliency values between sensitive and relevant features of penalized subjects at a conservative level of 1% (p-value). The obtained p-value of 0.0139, calculated using a two-sample t-test to compare accumulated mean sensitive and mean relevant values from the non-regularized model (see Tab. 1), was greater than the conservative threshold; thus returning no statistical evidence to reject the Null.
Please note that while Algorithm 1 establishes a linear correlation between relevant features and job classes (for each profile) and penalizes filtered subjects based on sensitive features, the bias-generating mechanism, applied here, does not provide statistical evidence of differing feature importance values when assigning labels to different job classes, according to the interpretation of the SM. This aspect offered a valuable scenario for our estimation: a situation in which bias perpetuates through neural processes, where sensitive and relevant features contribute equally to the unfair decision-making about penalized subjects. In other words, by assuming feature's equal contribution to unfair decisions, we reasonably estimated the amount of bias reduction as explained by the difference between mean relevant and mean sensitive feature values in profile matching the filter when applying regularization.

Following the above considerations, we conducted an ANOVA with post-hoc test to assess the effect of model type and estimate the amount of bias reduction provided by the regularized models. To this aim, we used the difference between mean relevant and mean sensitive feature values (referred to as the delta of saliency in this analysis) as the ANOVA dependent variable. Delta values were accumulated over folds for the models considered, and reported in Fig. 1. Only profiles that match the biased filter are included. ANOVA in Fig. 2 indicates a statistically significant effect of the model type on delta values ($p < 0.01$). Therefore, we conducted a post hoc test with the following results.



Non-Reg - Delta of Saliency Values (Profiles Matching Filter)

Fold	Mean Sensitive Saliency	Mean Relevant Saliency	Delta
1	0.052	0.273	0.221
2	0.019	0.053	0.035
3	0.02	0.043	0.023
4	0.104	0.234	0.131
5	0.032	0.069	0.037
6	0.054	0.152	0.098
7	0.031	0.063	0.032
8	0.096	0.143	0.047
9	0.065	0.203	0.138
10	0.075	0.109	0.034
Average	0.055	0.134	0.08

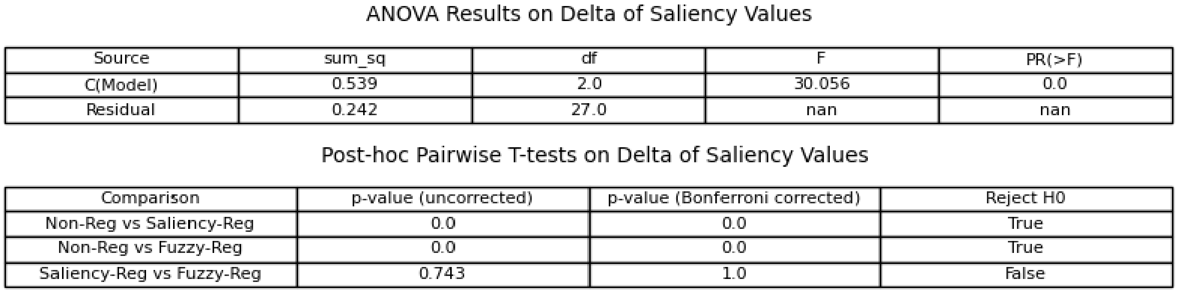
Saliency-Reg - Delta of Saliency Values (Profiles Matching Filter)

Fold	Mean Sensitive Saliency	Mean Relevant Saliency	Delta
1	0.002	0.191	0.189
2	0.01	0.476	0.465
3	0.018	0.441	0.423
4	0.01	0.428	0.418
5	0.001	0.431	0.431
6	0.0	0.414	0.414
7	0.011	0.275	0.265
8	0.002	0.375	0.374
9	0.002	0.333	0.331
10	0.009	0.415	0.406
Average	0.006	0.378	0.371

Fuzzy-Reg - Delta of Saliency Values (Profiles Matching Filter)

Fold	Mean Sensitive Saliency	Mean Relevant Saliency	Delta
1	0.034	0.283	0.249
2	0.038	0.264	0.226
3	0.102	0.579	0.477
4	0.012	0.554	0.542
5	0.066	0.487	0.421
6	0.048	0.5	0.452
7	0.058	0.454	0.396
8	0.029	0.222	0.193
9	0.048	0.273	0.225
10	0.023	0.399	0.376
Average	0.046	0.402	0.356

Figure 1: Delta of Saliency for the applied models.



ANOVA Results on Delta of Saliency Values

Source	sum_sq	df	F	PR(>F)
C(Model)	0.539	2.0	30.056	0.0
Residual	0.242	27.0	nan	nan

Post-hoc Pairwise T-tests on Delta of Saliency Values

Comparison	p-value (uncorrected)	p-value (Bonferroni corrected)	Reject H0
Non-Reg vs Saliency-Reg	0.0	0.0	True
Non-Reg vs Fuzzy-Reg	0.0	0.0	True
Saliency-Reg vs Fuzzy-Reg	0.743	1.0	False

Figure 2: Saliency Analysis: ANOVA with post-hoc test. Delta of Saliency is used as dependent variable of the ANOVA models.

- There is statistically significant difference in the delta of saliency values between Non-Reg vs Saliency-Reg (p = 0.0000).

- There is statistically significant difference in the delta of saliency values between Non-Reg vs Fuzzy-Reg (p = 0.0000).

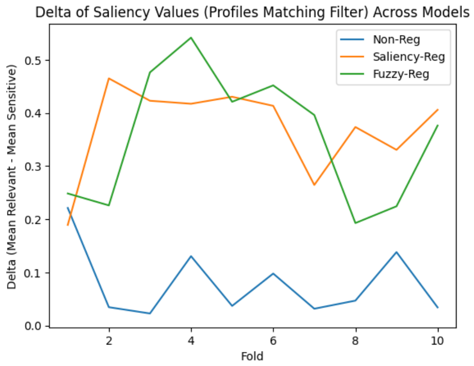


Figure 3: Delta of saliency across models for profiles matching the filters.

- There is no statistically significant difference in the delta of saliency values between Saliency-Reg vs Fuzzy-Reg (p = 1.0000).

Finally, we summarize these results qualitatively in Fig. 4 by reporting the bar-plot of the importance of each relevant and sensitive feature (Left column), for the neural decision making, averaged across validation folds (only profiles matching the filter are considered.)
These plots reveals, on average, how much each model relies on each sensitive and relevant feature when making predictions, thus highlighting again how the effect of regularization is distributed across the 2 groups of features.

3.2 Results

In conclusion, our experiments produced the following insights:

- **Bias Mitigation Effectiveness:** We estimated the "*effect*" of mitigation directly focusing on the difference between relevant and sensitive saliency, as explained by the saliency maps: higher delta values are observed when applying regularized models, thus implying higher mitigation capability concerning the non regularized models. No significant difference in delta is assessed between saliency and fuzzy-based regularization.
- **Consistency Across Folds:** The performance trend across (training vs. validation) folds (accuracy) provides a comprehensive view of each model's reliability and robustness under different data distributions (Fig. 4, right column)

These results demonstrate that in-processing regularization through the saliency maps and the integrated fuzzy logic allows mitigation of the predictive model's bias induced by data distortion. In particular, the mitigation achieved through regularization emphasizes the relationship between relevant and sensitive attributes, which is central to the decision-making process of this simulation study.

4 CONCLUSIONS

Fuzzy implications combined with saliency maps offer a promising strategy to make neural predictions more transparent and accountable while simultaneously addressing biases embedded in decision-making. In this paper, we have integrated a fuzzy system and saliency maps to penalize models that overly rely on sensitive features.

Experimental results demonstrate a significant reduction in the saliency values of sensitive features when fuzzy-based and saliency-based regularization are applied, thereby promoting fairness in decision-making. While no significant performance differences have been observed between fuzzy-based (integrated system) and saliency-based regularization, the integrated system offers additional advantages. It facilitates neural explainability (as per XAI principles) and enables the model to adapt pre-defined, human-understandable fairness constraints. This capability is particularly crucial in high-risk applications such as medical diagnostics.

It is important to emphasize that the numerical results presented here are affected by various factors that create degrees of freedom in the system. Elements like the choice of the Gaussian curve, its parameters, the rules formulated, and the fuzzy operators used for evaluations contribute to the complexity and variability of the outcomes. Therefore, future extensions will require experiments to constrain the system's degrees of freedom to better understand how these parameters influence the robustness of the performances obtained.

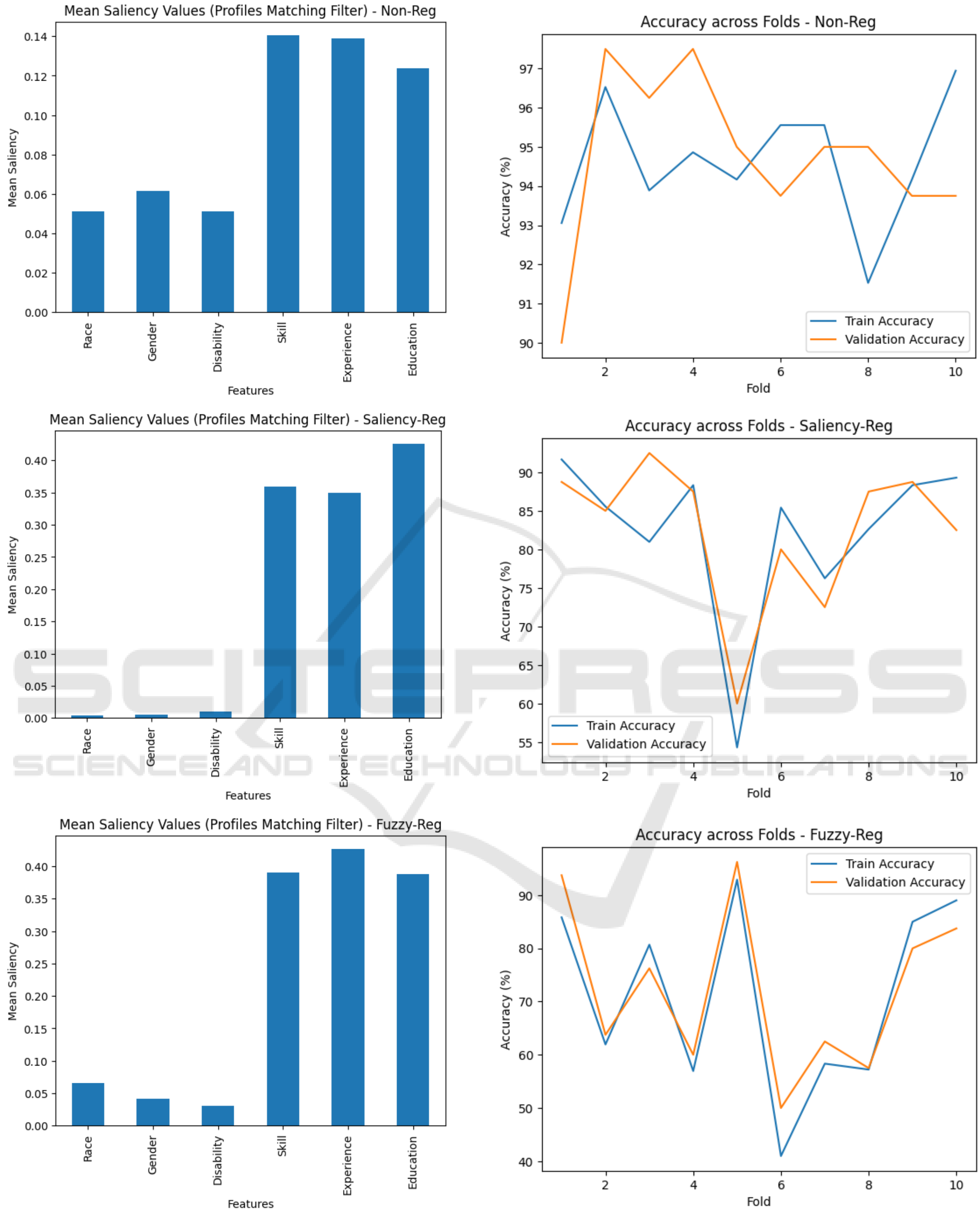


Figure 4: Saliency Regularization across feature (Left) and Train vs validation across folders (right).

ACKNOWLEDGEMENTS

This work was partially supported by the MUR under the grant “Dipartimenti di Eccellenza 2023-2027” of the Department of Informatics, Systems and Communication of the University of Milano-Bicocca, Italy.

REFERENCES

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. (2015). Certifying and removing disparate impact. *Proc. of the 21th ACM SIGKDD*, pages 259–268.

Ghosh, B., Basu, D., and Meel, K. S. (2023). ”how biased are your features?”: Computing fairness influence functions with global sensitivity analysis. In *Proc. of the 2023 ACM Conf. on Fairness, Accountability, and Transparency*, pages 138–148.

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

Iosifidis, V. and Ntoutsi, E. (2019). Adafair: Cumulative fairness adaptive boosting. In *Proc. of the 28th ACM international conference on information and knowledge management*, pages 781–790.

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

Kamiran, F., Calders, T., and Pechenizkiy, M. (2010). Discrimination aware decision tree learning. In *2010 IEEE conf. on data mining*, pages 869–874. IEEE.

Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J. (2012). Fairness-aware classifier with prejudice remover regularizer. In *Machine Learning and Knowledge Discovery in Databases: ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II 23*, pages 35–50. Springer.

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

Tjoa, E. and Guan, C. (2020). A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE trans. on neural networks and learning systems*, 32(11):4793–4813.

Zadeh, L. A. (1971). Similarity relations and fuzzy orderings. *Information sciences*, 3(2):177–200.

Zafar, M. B., Valera, I., Gomez-Rodriguez, M., and Gummadi, K. P. (2017). Fairness constraints: Mechanisms for fair classification. *Artificial Intelligence and Statistics*, pages 962–970.

Zhang, B. H., Lemoine, B., and Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. *Proc. of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340.