ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

# Improving Fairness via Intrinsic Plasticity in Echo State Networks

Andrea Ceni[1], Davide Bacciu[1], Valerio De Caro[1], Claudio Gallicchio[1], and Luca Oneto[2] *

1 - University of Pisa, Largo Bruno Pontecorvo 3 56127, Pisa, Italy

2 - University of Genoa, Via Opera Pia 11a, 16145, Genova, Italy

**Abstract**. Artificial Intelligence, and in particular Machine Learning, has become ubiquitous in today's society, both revolutionizing and impacting society as a whole. However, it can also lead to algorithmic bias and unfair results, especially when sensitive information is involved. This paper addresses the problem of algorithmic fairness in Machine Learning for temporal data, focusing on ensuring that sensitive time-dependent information does not unfairly influence the outcome of a classifier. In particular, we focus on a class of training-efficient recurrent neural models called Echo State Networks, and show, for the first time, how to leverage local unsupervised adaptation of the internal dynamics in order to build fairer classifiers. Experimental results on real-world problems from physiological sensor data demonstrate the potential of the proposal.

## 1 Introduction

Artificial Intelligence, and in particular Machine Learning (ML), is nowadays ubiquitous thanks to massive investments in making it a commodity. In some applications, e.g., games [1], healthcare [2], and text generation [3], these tools have been shown to compare to human capabilities. These achievements are accompanied by increasing concerns about their impact on society [4, 5]. In fact, real-world datasets often reflect historical biases in society and, when these data are fed to ML algorithms, they often result in models which actually exacerbate these biases [6]. In this paper, we deal with the problem of ensuring that ML models do not discriminate subgroups in the population based on, e.g., gender, race, or political and sexual orientation, namely to develop fairer ML models [7]. In particular, in this paper we focus on a specific notion of fairness called *(Difference of) Demographic Parity (DP)* [8]. DP is a fairness metric that measures to what degree machine learning model's predictions are independent of membership in a sensitive group. In other words, DP is maximised when the probability of a certain prediction is not dependent on sensitive group membership. For example, if we have two groups of people, 1 and 2, and we want to predict whether they will be approved for a loan or not, then the model achieves maximum DP when the same percentage of people from group 1 and 2 are approved for a loan. The problem of unfairness is present even with critical

*This work is partially supported by ELSA – European Lighthouse on Secure and Safe AI funded by the European Union under grant agreement No. 101070617, by project SERICS (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union - NextGenerationEU, by the EC H2020 programme under project TEACHING (grant n. 871385), and by EMERGE, a project funded by EU Horizon research and innovation programme (grant n. 101070918).

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

temporal data as in the area of healthcare [9]. This motivates us to focus on improving fairness on a class of deep learning models for sequential data under the paradigm of Reservoir Computing (RC). RC is a paradigm for training Recurrent Neural Networks (RNNs) particularly fast and energy-efficient. RC's key idea is to train only a readout layer (the actual classifier), while keeping untrained a large randomly connected hidden layer, called the reservoir. The reservoir is driven by the input time series, then the sequence of reservoir states is exploited to learn the readout layer, entitled of operating the actual classification. We make use of Echo State Networks (ESNs), a class of RC machines which owes its name to the fundamental assumption of stability of its internal neural dynamics, known in the literature as the echo state property. ESNs have been successfully used in time series classification tasks, and Intrinsic Plasticity (IP) has been shown to increase the accuracy of ESN classifiers building richer representations of the temporal features [10]. Here, we introduce the use of IP to efficiently pursue DP objectives by promoting the alignment of the reservoir activations' distributions, to make them indistinguishable on the basis of the fairness-critical information. We propose a modular hidden layer composed of multiple subreservoirs trained in a local and unsupervised fashion to blur potential discriminatory features in the data from the perspective of the classifier by architectural design. We test our approach by means of physiological sensor data in the area of human monitoring.

## 2 Improving Fairness in Reservoir Spaces

ESN models are efficient RNNs, belonging to the RC [11] category, which exploit the behavior of the recurrent layer as a discrete-time dynamical system. In ESNs, the recurrent layer is denoted as reservoir and is made up of a set of sparsely connected neurons. Given an input sequence of vectors $\mathbf{u}(t) \in \mathbb{R}^{N_U}$ with $t \in [T]$, the dynamics of a reservoir with $R$ leaky-integrator neurons is regulated by the following state transition function:

$$\mathbf{x}(t) = (1-a)\,\mathbf{x}(t-1) + a \tanh\Big(\mathbf{W}_{in}\mathbf{u}(t) + \mathbf{b}_{rec} + \hat{\mathbf{W}}\mathbf{x}(t-1)\Big) \qquad (1)$$

where $\mathbf{W}_{in} \in \mathbb{R}^{N_R \times N_U}$ is the input transformation matrix, $\hat{\mathbf{W}} \in \mathbb{R}^{N_R \times N_R}$ is the recurrent transformation matrix, $\mathbf{b}_{rec} \in \mathbb{R}^{N_R}$ is the reservoir bias term, and $\mathbf{x}(0) = \mathbf{0}$. The main advantage of using ESNs is that the matrices $\hat{\mathbf{W}}$ and $\mathbf{W}_{in}$ are kept fixed. This allows to avoid backpropagating the error signal through time and to learn the output transformation for the task at hand efficiently. In particular, a common choice is to address such learning problem as a linear system, and the output transformation $\mathbf{y}(t) = \mathbf{W}\mathbf{x}(t) + \mathbf{b}_{out}$ is learned by leveraging the closed-form equation of ridge regression, i.e., $\mathbf{W} = \mathbf{Y}\mathbf{S}^T(\mathbf{S}\mathbf{S}^T + \lambda\mathbf{I})^{-1}$, where $\mathbf{W} \in \mathbb{R}^{N_Y \times N_R}$ is the output transformation matrix, $\mathbf{Y} \in \mathbb{R}^{N_Y \times N_T}$ denotes the sequence of target labels, $\mathbf{S} \in \mathbb{R}^{N_R \times N_T}$ is the matrix of time-ordered reservoir states, $\lambda$ is an L2-regularization term, and $\mathbf{I}$ is the identity matrix.

In this paper, we propose ESNs that are *intrinsically fair*, meaning that, given a learning task where the input time series are coming from different sensitive groups (e.g., gender or ethnicity), the prediction is performed on temporal

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

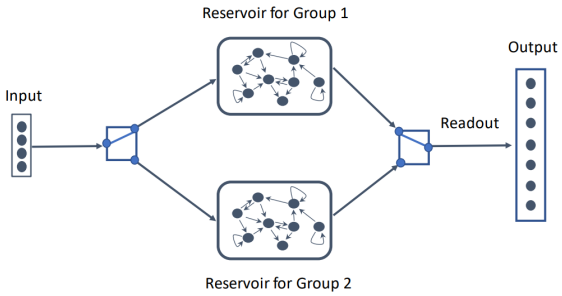


Fig. 1: Proposed RC architecture assuming the inputs are coming from two groups (1 and 2). Inputs are routed toward the *group-specific* reservoir, and the resulting sequence of states passes through a *common* readout layer.

features adapted to generate the same statistics regardless of the group membership. Without loss of generality, in Figure 1 we sketched an architecture for solving a learning task where the input data are coming from two subpopulations. Our architecture presents two reservoirs, namely $R_1$ and $R_2$, one for each group. Given an input sequence $\mathbf{u}$, $R_1$ is activated exclusively when $\mathbf{u}$ belongs to Group 1, and $R_2$ is activated exclusively when an input belongs to Group 2. Finally, the output of the ESN is computed by a readout layer common to both $R_1$ and $R_2$, meaning that it takes in input sequences of states from both reservoirs. Since the readout is trained by means of sequences that can come from either reservoir $R_i$, our aim is to avoid for its predictions to leverage on the subpopulations' biases. The fundamental assumption to achieve this objective is that *the dynamics of each group-specific reservoir must not expose the bias of the corresponding subpopulation*. To achieve this property, all the group-specific reservoirs are initialized with the same set of parameters (i.e., common $\hat{\mathbf{W}}$, $\mathbf{W}_{in}$, $b_{rec}$ and $\alpha$). Then, all of them are adapted *independently* via Intrinsic Plasticity (IP) [12], an unsupervised algorithm which improves the information gain of the reservoir. Formally, this algorithm reformulates the neurons' activation by including a gain and bias parameter to scale and translate the cumulative input of the neuron, i.e., $\tilde{x} = \tanh(gx_{net} + b)$. When using the tanh as non-linearity, the objective of IP is to minimize the Kullback-Leibler divergence between the empirical distribution of the neural activations and a desired Gaussian distribution with parameters $\mu$ and $\sigma$. Sparing the details of their derivation, the update rules applied by this algorithm are the following:

$$\Delta b = -\eta((-\mu/\sigma^2) + (\tilde{x}/\sigma^2 + 1 - \tilde{x}^2 + \mu\tilde{x})) \tag{2}$$

$$\Delta g = \eta/g + \Delta b x_{net} \tag{3}$$

where $\mu$ and $\sigma$ denote the mean and standard deviation of the target Gaussian distribution and $\eta$ is a learning rate. In our setup, each group-specific reservoir $R_i$ is equipped with a local set of gain and bias parameters $\mathbf{g}_i$ and $\mathbf{b}_i$, and the unsupervised adaptation of $R_i$ via IP is performed on the input data belonging to Group $i$ with a common target gaussian with parameters $\mu$ and $\sigma$. Besides

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

its common effect of improving the information gain and the performance of the ESN on the task at hand, adapting the group-specific reservoirs $R_i$ via IP allows for *having group-independent dynamics*. Informally speaking, the key intuition is that after the IP session, R1 and R2 are adapted to produce approximately the same Gaussian distribution of activations when driven with the inputs belonging to, respectively, Group 1 and Group 2. This makes it harder for the subsequent predictor to exploit any discriminatory feature based on the group membership to increase the accuracy. A similar idea has also been leveraged in [13], showing its effectiveness in the context of end-to-end trainable deep feed-forward models. Finally, note that, to leverage this approach, the sensitive feature need to be available in the testing phase which might not always be possible to due to legal requirements [14].

## 3 Experiments

We test our methodology on two time series classification datasets: WESAD [15], and the Tufts fNIRS Mental Workload (TfMW) [16]. WESAD is a multimodal dataset for stress and affect detection from both a wrist- and a chest-worn device, which was collected from 15 participants in a ≈36-minute session where they performed activities depending on the cognitive state to be induced. Specifically, we train binary classifiers focusing on the two classes "stress" and "amusement" of the WESAD dataset, and measure the fairness with regard to the two populations of males and females. TfMW is a dataset composed of ≈30-seconds of multivariate recordings from a sensor probe placed on the forehead. These time series representing brain activity throughout the session are used as input for the prediction of the mental workload intensity of the user during that window of time. In particular, we use only TfMW data from subjects declared as either white or asian, resulting in a total number of 60 subjects. Specifically, we train binary classifiers focusing on the two classes of workload intensity "0-back" and "2-back" of the TfMW dataset, and measure the fairness with regard to the two populations of whites and asians. We make experiments on 2 models. The first is eq. (1), without the use of IP. The second is our proposed model described in Section 2 and depicted in Figure 1. For both models, we use reservoirs of 100 neurons. The IP algorithm is always set with a mean value of $\mu = 0$, and run for a fixed number of 10 epochs to fix the computational budget. We perform validation via a random search on the following grid of hyperparameters: $\rho \in \{0.5, 0.7, 0.9, 1.1\}$, $\omega \in \{0.1, 0.5, 1., 1.5\}$, $a \in \{0.01, 0.1, 1\}$, $\lambda \in \{0, 0.001, 0.1\}$, for all models, and also $\sigma \in \{0.01, 0.1, 1\}$, $\eta \in \{0.001, 0.01, 0.1\}$ for the model with IP. The hyperparameters $\rho$, and $\omega$, rescale the reservoir matrix $\hat{\mathbf{W}}$, and the input-to-reservoir connections $\mathbf{W}_{in}$ of eq. (1), respectively. While $a$, $\lambda$, $\sigma$, and $\eta$, are defined in Section 2. Note that the possible combinations of hyperparameters for the model with IP are 6 times larger than the one without IP. Therefore, to ensure a similar exploration of the hyperparmater's grid, the model with IP has been run with 6 times more random trials. The WESAD dataset is composed of 15 subjects in total. We perform leave-one-out double cross-validation. For each subject left out as test, use remaining 14 subjects for training. Specifically, we use 10 subjects for actual

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

| Dataset | WESAD | TfMW |
|---|---|---|
| no-IP acc | 71.97% | 56.92% |
| IP acc | 78.16% | 58.56% |
| no-IP fair | 93.11% | 95.03% |
| IP fair | 97.14% | 96.98% |

Table 1: Test set accuracy and fairness measure on WESAD and TfMW, the subpopulations encompass gender and ethnicity biases respectively.

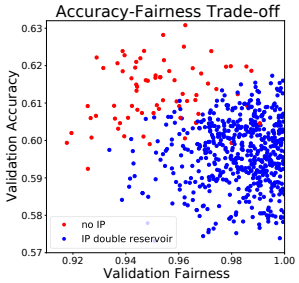


Fig. 2: Accuracy-Fairness trade-off on validation data of TfMW without IP (red) and with IP (blue).

training, and 4 subjects for validation. We end up with a set of hyperparameter to test on each of the 15 subjects. Therefore, for each test subject left out, we train on the remaining 14 training subjects with the hypeparameter combination yielding the best validation accuracy, and compute test accuracy and fairness DP metric. Then, we provide the final test accuracy as the arithmetic mean over all the subjects. The subset of TfMW dataset that we consider in our experiments is composed of 65 subjects in total. We shuffle the 65 subjects and partition them into 13 buckets of 5 subjects each. We perform a similar leave-one-out double cross validation as we did on the WESAD dataset, but here on a per bucket basis. Specifically, for each bucket left out as test, we use the remaining 12 buckets for training, 9 buckets for actual training, and 3 buckets for validation. The results of the validation session on the TfMW dataset are reported in Figure 2. From Figure 2 emerges a trade-off between accuracy and fairness, in line with previous works in the literature [17, 18]. When randomly selecting hyperparameters, the model with IP tends to have on average larger values of fairness, at the price of slightly decreasing the accuracy. Interestingly, when performing model selection based on the maximisation of the accuracy on validation, this trade-off disappears. As shown in Table 1, the final results of accuracy and fairness on the test session reveal an increase in both the metrics of accuracy and fairness. While the increase of accuracy on test might be attributed solely on the better representations of the temporal information built by IP, the increase of fairness is due to the intrinsic architectural bias provided by our proposed model. These experiments confirm the validity of our methodology. The discriminatory biases contained in the data get blurred into each group-specific reservoir making harder for the subsequent readout classifier to leverage on discriminatory features to increase the accuracy.

## 4 Conclusions

In this paper, we proposed an RC-based model to perform classification with the goal of improving the fairness of the classifier. Our methodology leverages

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

on the IP algorithm, creating different reservoirs that produce approximately similar features when fed with data from different subpopulations. This makes it harder for the classifier to rely on discriminative features to increase accuracy. From experiments based on real physiological data, we have shown that our proposed model can improve fairness while achieving higher accuracy than the baseline case without IP. The already promising results reported in this paper motivate us to make efforts to generalize the proposed methodology to a single reservoir configuration, rather than a modular one composed of sub-reservoirs. Future work will also explore forms of local unsupervised reservoir adaptation that explicitly address the fairness metrics. Future work will also investigate extensions of the introduced approach to different deep learning architectures, e.g. convolutional neural networks and large language models.

## References

[1] D. Silver, A. Huang, C. J. Maddison, and Others. Mastering the game of go with deep neural networks and tree search. *Nature*, 529:484–489, 2016.

[2] J. De Fauw, J. R. Ledsam, B. Romera-Paredes, and Others. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine*, 24:1342–1350, 2018.

[3] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[4] L. Floridi. Establishing the rules for building trustworthy ai. *Nature Machine Intelligence*, 1:261–262, 2019.

[5] W. Liang, G. A. Tadesse, D. Ho, and Others. Advances, challenges and opportunities in creating data for trustworthy ai. *Nature Machine Intelligence*, 4:669–677, 2022.

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

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

[8] D. Franco, L. Oneto, and D. Anguita. Fair empirical risk minimization revised. In *International Work-Conference on Artificial and Natural Neural Networks*, 2023.

[9] K. Bhanot, M. Qi, J. S. Erickson, I. Guyon, and K. P. Bennett. The problem of fairness in synthetic healthcare data. *Entropy*, 23(9):1165, 2021.

[10] X. Wang, Y. Jin, and K. Hao. Synergies between synaptic and intrinsic plasticity in echo state networks. *Neurocomputing*, 432:32–43, 2021.

[11] M. Lukoševičius and H. Jaeger. Reservoir computing approaches to recurrent neural network training. *Computer Science Review*, 3(3):127–149, 2009.

[12] B. Schrauwen, M. Wardermann, D. Verstraeten, J. J. Steil, and D. Stroobandt. Improving reservoirs using intrinsic plasticity. *Neurocomputing*, 71(7-9):1159–1171, 2008.

[13] L. Oneto, M. Donini, G. Luise, C. Ciliberto, A. Maurer, and M. Pontil. Exploiting mmd and sinkhorn divergences for fair and transferable representation learning. In *Neural Information Processing Systems*, 2020.

[14] C. Dwork, N. Immorlica, A. T. Kalai, and M. D. M. Leiserson. Decoupled classifiers for group-fair and efficient machine learning. In *Conference on Fairness, Accountability and Transparency*, 2018.

[15] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *ACM international conference on multimodal interaction*, 2018.

[16] Z. Huang, L. Wang, G. Blaney, and Others. The tufts fnirs mental workload dataset & benchmark for brain-computer interfaces that generalize. In *Neural information processing systems*, 2021.

[17] I. Chen, F. D. Johansson, and D. Sontag. Why is my classifier discriminatory? In *Neural information processing systems*, 2018.

[18] A. K. Menon and R. C. Williamson. The cost of fairness in binary classification. In *Conference on Fairness, accountability and transparency*, 2018.