# Deep Neural Network-Aided Soft-Demapping in Coherent Optical Systems: Regression Versus Classification

Pedro J. Freire, *Member, IEEE*, Jaroslaw E. Prilepsky, Yevhenii Osadchuk, *Member, IEEE*,
Sergei K. Turitsyn, *Senior Member, IEEE*, and Vahid Aref

***Abstract*— We examine here what type of predictive modelling, classification, or regression, using neural networks (NN), fits better the task of soft-demapping based post-processing in coherent optical communications, where the transmission channel is nonlinear and dispersive. For the first time, we present possible drawbacks in using each type of predictive task in a machine learning context, considering the nonlinear coherent optical channel equalization/soft-demapping problem. We study two types of equalizers based on the feed-forward and recurrent NNs, for several transmission scenarios, in linear and nonlinear regimes of the optical channel. We point out that even though from the information theory perspective the cross-entropy loss (classification) is the most suitable option for our problem, the NN models based on the cross-entropy loss function can severely suffer from learning problems. The latter translates into the fact that regression-based learning is typically superior in terms of delivering higher Q-factor and achievable information rates. In short, we show by empirical evidence that loss functions based on cross-entropy may not be necessarily the most suitable option for training communication systems in practical scenarios when overfitting- and vanishing gradients-related problems come into play.**

***Index Terms*— Neural networks, nonlinear equalizer, classification, regression, coherent detection, digital signal processing, optical communications.**

## I. Introduction

TO IMPROVE the performance of optical fiber systems, it is important to mitigate the detrimental impact of linear and, most importantly, nonlinear transmission impairments that cape the systems' throughput [1], [2], [3]. Numerous digital signal processing (DSP) algorithms have been proposed and studied for the optical fiber channel equalization problem [3]. Over the past few years, the "conventional" equalizers/soft-demappers have started to evolve toward the designs incorporating machine learning techniques [4], [5], [6]. In particular, the neural network (NN) based channel equalization/demapping has recently become a topic of intensive research in optical communications, due to its capability in mitigation of linear and nonlinear impairments in optical channels and transceivers. For instance, in [7], [8], [9], and [10] [7], [8], [9], [10] the authors successfully applied artificial NN-based nonlinear equalizers for impairment compensation in coherent transmission links, while in [11] and [12], more advanced deep learning algorithms were introduced and compared with the performance rendered by conventional DSP methods. Additionally, in [13] the authors suggested a bit-wise soft-demapper model based on bidirectional recurrent NN applied for nonlinear ISI compensation in coherent links. Recurrent NNs in the form of long short-term memory (LSTM) has shown great potential in nonlinearities mitigation [14], [15]. Additionally, the end-to-end learning designs [16] that deal with coherent constellation optimization have been intensively studied in the literature [17], [18], [19], [20], [21], [22], [23]: in these works, the end-to-end system is typically composed as an auto-encoder trained with the cross-entropy loss (CEL) function. Finally, we notice that the authors of [24] recently investigated the back-propagation blocking problem in training the end-to-end NN designs, showing the possible problems in using the CEL. In the same [24], a version of the mean squared error (MSE) loss, for the end-to-end systems, was proposed, where the resulting performance was better than the one delivered by the CEL-based NN models.

The most popular and, perhaps, simple approach to improving the channel capacity is the receiver-based equalizer – a special-purpose DSP device that can (partially) reverse the distortions incurred by a signal when passing over the optical channel. These equalizers are typically designed and optimized based on the minimum-mean-squared-error (MMSE) criteria. The usefulness of MMSE equalizers is stipulated for several reasons, including: (i) the MMSE is quite convenient for mathematical optimization because of convexity and differentiability of its objective function; and (ii) the MMSE equalizer is

Manuscript received 18 October 2021; revised 7 March 2022, 8 June 2022, and 21 August 2022; accepted 21 September 2022. Date of publication 10 October 2022; date of current version 19 December 2022. This paper was supported by the EU Horizon 2020 program under the Marie Sklodowska-Curie grant agreement 813144 (REAL-NET). Jaroslaw E. Prilepsky is supported by Leverhulme Trust, Grant No. RP-2018-063. Sergei K. Turitsyn acknowledges support of the EPSRC project TRANSNET. The associate editor coordinating the review of this article and approving it for publication was L. Schmalen. *(Corresponding author: Pedro J. Freire.)*

Pedro J. Freire, Jaroslaw E. Prilepsky, Yevhenii Osadchuk, and Sergei K. Turitsyn are with the Aston Institute of Photonic Technologies, Aston University, B4 7ET Birmingham, U.K. (e-mail: p.freiredecarvalhosouza@aston.ac.uk).

Vahid Aref is with the Nokia Bell Laboratories, 70435 Stuttgart, Germany (e-mail: vahid.aref@nokia.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCOMM.2022.3213284.

Digital Object Identifier 10.1109/TCOMM.2022.3213284

usually optimized independently of the underlying waveform or modulation format. In this paper, we call this category of NN models based on MMSE as a regression equalizer (Reg.), but we can term them as a dual-stage soft-demapping (i.e. the NN MMSE post-equalizer followed by the AWGN demapper) [25].

Another approach is to design a model incorporating the decision step, which corresponds to the classifier in the machine learning literature, see, e.g., [26, Chapter 6]. Such a device is convenient insofar as the transmitted signal in digital communications is usually generated from a discrete finite-size constellation, e.g. quadrature amplitude modulation, QAM. This approach has received more attention recently [13], [16], [23], [27], [28], [29], in view of the following reasons: (i) the classifier is optimized for the specific in-use modulation format; (ii) it directly maximizes the information rate, the main objective of the channel equalization, and outputs the likelihoods for each received symbol, a more suitable metric for the subsequent forward error correction; (iii) and, even more importantly, it can adapt itself to the correct statistical channel characteristics. In our paper, we call this category of predictive model a multi-class classifier (MC Class.), and in our communication study case, we can see them as single-stage soft demapping, in which we use the categorical CEL in the learning process.

Here, we also mention that another possible approach allows us to build on the bitwise representation of the received symbols and to train the NN model using the binary cross-entropy loss (BCEL) as a one-step setup [29]. This type of receiver model, named in machine learning as a *multi-label classifier* (MB Class.), is used in end-to-end learning [17], [23], [24] or for the soft-demapping in short-reach transmission scenarios [29]. In our current work, we mostly focus on the multi-class classification and the regression approaches, but in the last part of the results section, we additionally perform a comparative study for the multi-label classification, to make sure that the observed effects refer to this case as well. The NN models for the regression equalizer, multi-class classifier, and the multi-label classifier used in this paper are schematically given in Fig.1.[1]

It is worth mentioning at this point that the transfer of the methods developed in the field of machine learning to optical communications should be taken into account *the underlying peculiarities and challenges of NN algorithms themselves*, while those are often overlooked when designing the equalizers/soft-demappers in communications-related literature. From [25] we infer that cross-entropy is the most suitable loss function for communication applications, accounting for its meaning in information theory. However, in our paper, we intend to increase the researcher's awareness of possible training problems that we observed when dealing with the CEL gradient-descent-based learning. Our case study here specifically refers to the complex problem of symbols demapping in long-haul coherent optical transmission, in which the memory from dispersion and fiber nonlinearity plays an important role; however, this may have less relevance to the short-haul/back-to-back scenarios, which, perhaps, deserve a separate study.

A fair comparison between the regression and classification is quite challenging, as these two produce different output variable types: discrete versus continuous, respectively. Such comparison studies have been carried out in theoretical machine learning-related works, where the output of the classification was analyzed with different loss functions [30], [31], [32]. However, to our knowledge, only in [33] directly explained the motivation to solve problems with regression instead of classification for specific tasks. In [34], [35], [36], [37], [38], and [39], it was recognized that both regression and classification tasks have potential downsides: the regression model cannot utilize the full flexibility of discriminative NN models (meaning that the statistic of the regression model is assumed to be Gaussian, while the classification model does not depend on the type of statistics, and, thus, is more flexible.); the classification model, in turn, is not able to capture misclassification differences when we misinterpret the classes, which can degrade the modelling quality. Thus, in these works, the combination of regression and classification was proposed as the best problem fit, but we do not address it here.

The interest in finding the best solution between the regression- and classification-based NN models, frequently arising in different areas, prompts us to conduct the investigation of this dilemma for the development of NN-based soft-demappers in coherent optical systems. For the optical channel demapping, this comparison may be made more evident by contrasting the multi-class classifier output to that obtained with the regression, in terms of bit error rate (BER) (i.e. using a hard decision metric) or with respect to the achievable information rate (AIR), i.e. using a soft decision metric. In this paper, we compare the performance of the multi-class CEL classifier and regression MMSE predictive models and expose the potential drawbacks of each task specifically for the NN-based long-haul optical channel soft-demapping problem. Our findings and rationale behind the observed interplay between the regression and classification results can be briefly summarized as follows.

- For the MSE training process we recommend using an *early stopping at the point where the achievable information rate is maximal*, but not at the point of minimal MSE, as to avoid underestimation in the calculation of the achievable information rate and "jail-window" constellation patterning [40].
- The CEL landscape is prone to sharp local minima, revealing large gradients at the points where the training loss value is close to zero. Thus, we arrive at the overfitting of the model, where the loss value for the training is much lower than that for the testing dataset. In contrast, the MSE produces wide local minima, where we have relatively small gradients in the vicinity of the training loss minimum. It allows us to avoid overfitting more efficiently.
- The CEL was incapable of delivering sufficient training quality due to the vanishing of back-propagating

[1] To get the bits out of the probabilities in the multi-label classification case, we use a simple decision block that returns 1 if the probability was higher than 0.5 and 0 otherwise.

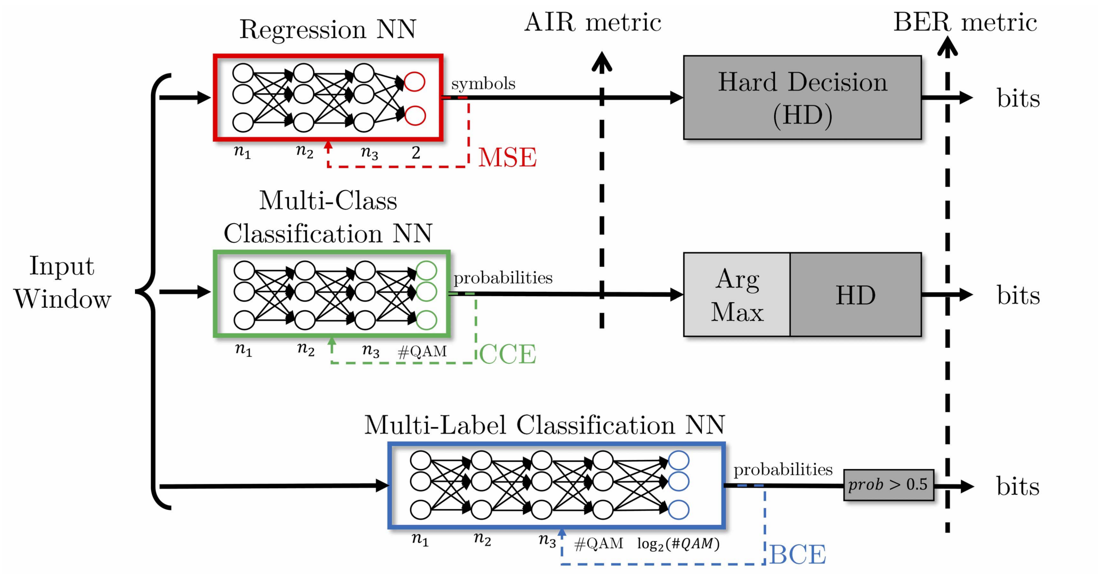


Fig. 1. Scheme of the three classic configurations of NN models, using the MLP as an exemplary NN core: regression (top), multi-class classification (middle), and multi-label classification (bottom), in the context of channel equalization and soft demapping in communications.

gradients. In contrast to regression (with the MSE loss function), classification could not handle well *the high-accuracy systems* (with a very low number of errors in the training datasets) that we typically have in optical communication applications, causing the NN to rapidly converge to a local minimum, where the gradient is close to zero.

- As it was also observed in computer vision tasks [41], the CEL function in our case gets overwhelmed by the "easy prediction", which is typical for the high accuracy systems.

These effects, and their consequence on the NN's learning quality and performance, are considered in detail further. As a result, it is exactly the classifiers' training challenges due to the aforementioned reasons, but not the optimally/suboptimality of the loss function for the demapping, that plays a crucial role in defining the performance of NN-based models. This conclusion can be somewhat counter-intuitive in view of the information theory arguments based on the optimality of the demapper [25]. we also point out that our findings here coincide with those presented in very recent [24] and [32], allowing us to conclude that the observed effects can pertain not only to the NN-based equalization/demapping but also to the other types of NN-based systems (e.g. to end-to-end learning).

## II. Deep Neural Network-Aided Soft-Demapping

### A. Regression-Based Equalizers and Multi-Class Classifiers

Equalization is the task of recovering the transmitted data $X_n$ from the received data $Y_n$. It maps $Y_n$ to the most likely transmitted data, $\hat{X}_n = f(Y_n;\Theta)$, for a given mapping function $f(\cdot)$ and the set of trainable parameters $\Theta$, optimized according to some likelihood measure. In our case of NN-based equalization, $f(\cdot;\Theta)$ denotes the NN itself, and $\Theta$ denotes its trainable weights and biases. Here, $X_n$ and $Y_n$ can denote either a single sample of transmitted and received data or the sequences of samples for either one or both of them (see the explicit explanations of our NN structure below in Subsec. II-B). For simplicity of presentation, we assume the single sample representation, and that $X_n$ is chosen from a constellation alphabet $\{c_1, c_2, \ldots, c_m\}$ with $c_i \in \mathbb{C}$, the complex space.

In regression-based equalization, $f(Y_n;\Theta)$ is relaxed to $f_{\text{reg}}(Y_n;\Theta)$ outputting any complex value, and its likelihood maximization boils down to the minimization of MSE, i.e. to finding the specific set of parameters $\Theta^*_{\text{reg}}$:

$$\Theta^*_{\text{reg}} = \text{argmin}_\Theta \left\{\mathbb{E}_{X_n,Y_n}\left[|X_n - f_{\text{reg}}(Y_n;\Theta)|^2\right]\right\}. \tag{1}$$

The above expectation $\mathbb{E}_{X_n,Y_n}$ is taken over the samples of transmitted data and the corresponding received data. Those samples are, in fact, distributed according to $P(X_n, Y_n) = P(X_n)P(Y_n|X_n)$, where $P(X_n)$ is the transmitted signal distribution and $P(Y_n|X_n)$ describes how likely the channel output $Y_n$ is upon the transmission of $X_n$.

In the classification-based equalization, i.e. effectively combining the regression with soft demapping into a single NN, we have: $\hat{X}_n \in \{c_1, c_2, \ldots, c_m\}$. In this case, $f(Y_n;\Theta)$ is relaxed to $f_{\text{cl}}(Y_n;\Theta)$ outputting a vector of posterior probabilities $(q_1, \ldots, q_m)$, where $q_k := Q(X_n = c_k|Y_n;\Theta)$, showing how likely $X_n = c_k$ are, given receiving $Y_n$. Then, $\hat{X}_n$ is equal to the $c_k$ that has the largest posterior probability. It turns out that the "maximum likelihood" estimation (the best effort of the model $f_{\text{cl}}(\cdot)$) is obtained if the following categorical CEL of the actual posterior $P(X_n|Y_n)$ and $Q(X_n|Y_n;\Theta)$ is minimized:

$$\mathcal{X}(P,Q;\Theta) = -\mathbb{E}_{Y_n}\left[\sum_{k=1}^{m} P(c_k|Y_n)\log_2(Q(c_k|Y_n;\Theta))\right]. \tag{2}$$

Equivalently, one can instead maximize

$$I_\Theta(X_n;\hat{X}_n) = \mathbb{E}_X[\log_2(P(X_n))] - \mathcal{X}(P,Q;\Theta), \tag{3}$$

where $I_\Theta(X_n;\hat{X}_n)$ is the achievable information rate for the mismatched decoding rule $Q(X_n|Y_n;\Theta)$ [42, Def. 12], [43, Theorem 2]. As a result, the classification-based equalization is optimized for the following set of parameters:

$$\Theta^*_{\text{cl}} = \text{argmax}_\Theta \left\{ I_\Theta(X_n;\hat{X}_n) \right\}. \tag{4}$$

Note that $I_\Theta(X_n;\hat{X}_n) \leq I(X;Y)$, the true mutual information of the channel, and the equality holds if $Q(X_n|Y_n;\Theta^*_{\text{cl}}) = P(X_n|Y_n)$ for all $(X_n,Y_n)$.

In the case of regression-based equalization, the AIR cannot be expressed directly from the optimization cost, Eq. (1). Instead, we computed an AIR using the Gaussian approximation of conditional probabilities (a mismatched distribution) by

$$\tilde{Q}(\hat{X}_n|X_n = c_k) = \frac{1}{2\pi\sqrt{\det(\Sigma_k)}} \exp\left(-\frac{1}{2} Z_{n,k}\Sigma_k^{-1} Z_{n,k}^T\right),$$

where $Z_{n,k} = (\text{Re}\{\hat{X}_n - \mu_k\}, \text{Im}\{\hat{X}_n - \mu_k\})$ and $\Sigma_k = \mathbb{E}[Z_{n,k}^T Z_{n,k}|X_n = c_k]$, with trainable mean $\mu_k$ and covariance matrix $\Sigma_k$. Correspondingly, we define

$$\tilde{I}_{\text{reg}}(X_n;\hat{X}_n) = \mathbb{E}_{X_n,Y_n}\left[\log_2\left(\frac{\tilde{Q}(\hat{X}_n|X_n)}{\sum_{k=1}^m P(c_k)\tilde{Q}(\hat{X}_n|c_k)}\right)\right]. \tag{5}$$

We use this achievable information rate to compare with the classification-based equalizer in (3). However, one should take into account that $\tilde{I}_{\text{reg}}(X_n;\hat{X}_n)$ *underestimates the true mutual information* $I(X_n;\hat{X}_n)$ [42], [44].

Note that it is known that the regression-based equalization in (1) can be expressed as a classification-based one in (4) with some special $f_{\text{cl}}(Y_n;\Theta)$ outputting posterior probabilities based on a complex Gaussian distribution:

$$Q_{\text{G}}(X_n|Y_n) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{|X_n - f_{\text{reg}}(Y_n;\Theta)|^2}{2\sigma^2}\right), \tag{6}$$

with some variance of $\sigma^2$ for each real and imaginary tributary. Accordingly,

$$\log_2(Q_{\text{G}}(X_n|Y_n)) = -\frac{|X_n - f_{\text{reg}}(Y_n;\Theta)|^2}{2\ln(2)\sigma^2} - \log_2(2\pi\sigma^2), \tag{7}$$

and

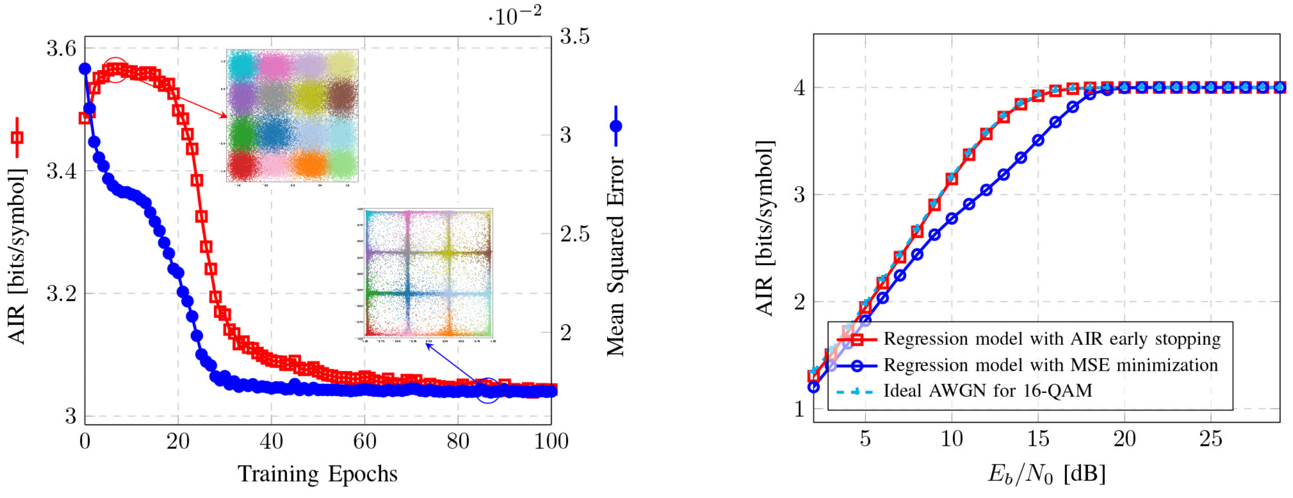
$$\begin{aligned}\mathcal{X}(P,Q_{\text{G}};\Theta) = \; &\frac{1}{2\ln(2)\sigma^2}\mathbb{E}_{X_n,Y_n}\left[|X_n - f_{\text{reg}}(Y_n;\Theta)|^2\right] \\ &+ \log_2(2\pi\sigma^2).\end{aligned} \tag{8}$$

One can verify that $\mathcal{X}(P,Q_{\text{G}};\Theta)$ is minimized if $\Theta = \Theta^*_{\text{reg}}$ and we choose $\sigma^2_{\text{reg}} = \frac{1}{2}\mathbb{E}_{X_n,Y_n}\left[|X_n - f_{\text{reg}}(Y_n;\Theta^*_{\text{reg}})|^2\right]$, the MSE. However, we need to point out one important pitfall already reported in [40]. Regarding the quality of transmission metrics, when performing the regression task via the MSE, the equalization model, if not monitored, starts to distort the QAM constellation into a so-called "jail window" pattern. The consequence of this is that the AIR calculation of such a new constellation becomes underestimated if using a Gaussian demapper with $\sigma^2_{\text{reg}}$. At this point, two things can be done to solve this problem: i) we can use an optimizer (e.g. linear search algorithm) to find a new $\sigma^2$ that maximizes the achievable information rate or even embed in the original NN the demapping process to learn this step jointly using the CEL loss function; ii) we can monitor the AIR of the model for a validation dataset and apply early stopping considering not the MSE loss function, *but the AIR metric itself.*[2] Solution ii) is simpler and produces an NN architecture that is less computationally complex compared to solution i).

To illustrate the effectiveness of solution ii) and the impact of the "jail window" constellation, we trained the MLP equalizer on a pure AWGN channel. The codes for this test with specific details are available in [46]. Fig. 2 summarizes the analyses done for this first investigation. In Fig. 2 (a), the MLP model is trained with a bit-energy to noise power ratio ($E_b/N_0$) equal to 12 dB, and a validation dataset is used to monitor both the AIR and MSE over 100 epochs. As can be seen, initially the AIR increases and the MSE decreases until a certain epoch when the AIR and MSE become almost constant. However, when we continue to train the NN, the MLP equalizer finds a set of weights that can further decrease the MSE, but this does not result in any additional gain in the AIR, but the AIR tendency is quite opposite because now the solution produces the "jail window" constellation. We have highlighted in Fig. 2 (a) the constellations when the model is saved on the highest validation AIR, and when the model is saved on the lowest validation MSE. To show the consequence of such a wrong early stopping in the testing performance, we have tested the model that we trained at $E_b/N_0 = 12$ dB, with $E_b/N_0$ varying from 2 dB to 29 dB, where we used the weights for the AIR early stopping and the "pure" MSE minimization. Fig. 2 (b) shows the result of our scrutiny, where we can see that using the model with the MSE minimization and $\sigma^2_{\text{reg}}$ for demapping produced the worst AIR value in the testing dataset than the situation when we used the model with the AIR early stopping and the same $\sigma^2_{\text{reg}}$ for the demapping. Option ii) is the method utilised by the regression models in this paper, and as seen in the Fig. 2 (b), it reached the same AIR as the ideal AWGN (Shannon limit) for 16 QAM.

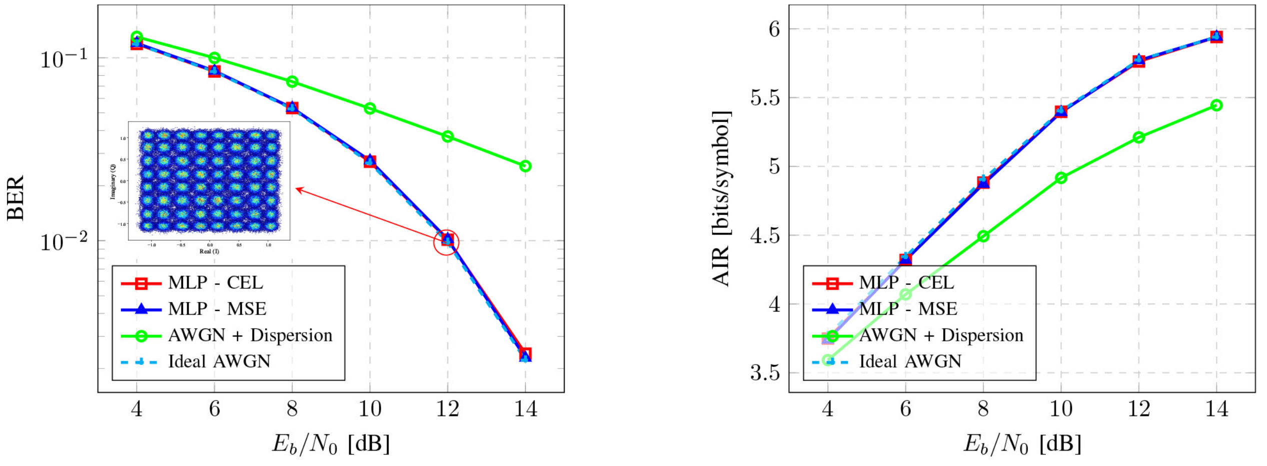
To show further the empirical confirmation that when using the AIR early stooping, the performance of the regression model does not get reduced compared to option i), where we have a NN model that learns the demapping jointly, we conduct the following investigation. We have simulated the simple 32 GBd, 64QAM transmission, where 3 km of linear fiber with the second-order dispersion parameter $D =$ 17 ps/(nm·km) is included to add some small amount of linear dispersion (losses and nonlinearities are not considered) and after the fiber, an AWGN is added with $E_b/N_0$ varying from 4 dB to 14 dB. At the receiver, we consider a pulse-shaping filter, and we have the same NN-based MLP as described in Section II-B, operating with 1 sample/symbol, to recover the dispersion introduced by the fiber. With this experiment,

[2] Another alternative, recently proposed in [45], can also be used to solve the "jail window" problem: it suggests incorporating an entropy regularization into the MSE function. This loss is referred to as the MSE-X loss function.



(a) (b)

Fig. 2. MLP equalizer based on the regression (MSE) for a pure AWGN channel with a 16QAM constellation. (a) represents the AIR and MSE metrics for a validation dataset over the training phase; (b) shows the AIR of a testing dataset when using the trained MLP model in (a) when using the weights of the epoch that produced maximum validation AIR (AIR early stopping) or when using the weights that produced minimum validation MSE (standard MSE minimization).



(a) (b)

Fig. 3. MLP NN based on the regression (MSE) and multi-class classifier (CCEL) investigation over the BER (a) and AIR (b) metrics for the AWGN system with 3 km linear fiber dispersion with 64 QAM. The constellation for validation dataset after the MLP-MSE equalizer with AIR early stopping is presented for the toy case when $E_b/N_0 = 12$ dB.

we want to show that the NN equalization based on the MSE with AIR early stopping and the NN based on the CEL will provide the same AIR and BER as we have for the system with only the AWGN and no fiber (we mark the latter as "Ideal AWGN").

In Fig. 3, the best values of AIR and BER achieved by the NN equalizers/soft-demapping based on either the MSE or the CEL are presented for different levels of $E_b/N_0$, when using the same NN training procedure described in Sec II.D. We also plotted the equalized constellation histograms after the NN based on the MSE for an exemplary point $E_b/N_0 = 12$ dB using a testing dataset. This constellation shows that the symbols after the NN based on MSE retain the form of the AWGN-induced constellation points histogram. As expected, both loss functions led to learning the linear chromatic dispersion filter quite precisely, and both NNs reached the performance of the "Ideal AWGN" system. This empirical example shows that a simple linear problem governed by AWGN can be learned using either the MSE with AIR early stopping or the CEL.

At the same time, the situation in which we consider a long nonlinear fiber transmission differs from the above simple case. Since we use an optimization algorithm based on gradient descent, the training can also *suffer from many NN-learning-related pitfalls*. Therefore, the goal of our paper is to show that this statistical limitation/mismatch of the MSE in regression can be "less harmful" to the AIR (or any eventual quality metric used) than the gradient learning-related problems that the CEL can (and typically does) produce in the case of classification. In short, our main result is that

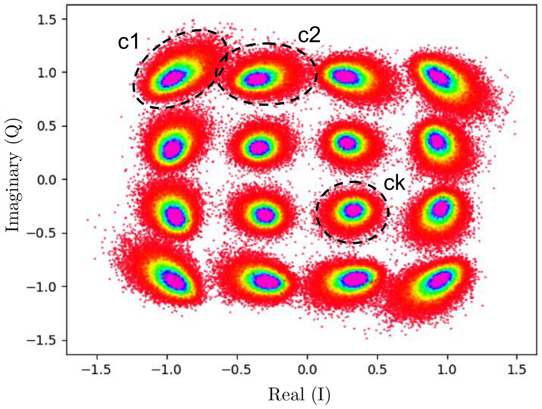


Fig. 4. Distribution of the received symbols in an ideal transmitter of SC-DP 16QAM case with launch power $-1$ dBm and 34.4 GBd, over $9 \times 50$ km TWC fiber. The constellation is given after a standard Rx-DSP described in Sec III.

the NN learning/training-related problems typically downgrade the ultimate quality metrics more than the MSE distribution-related mismatch, insofar as it is virtually impossible to achieve the "ideal" training in the CEL classification case in reality, and the resulting penalty "always wins" (again, in contrast to the simple 3 km linear fiber case presented above).

To make some general notes, as mentioned previously, when AWGN is the primary source of signal distortion the regression-based linear equalization with MSE loss, Eq. (1), can be quite an appropriate choice in terms of training [25]. However, such equalizers are penalized if the distortion statistics deviate from the AWGN statistics. This typically occurs in optical communication, as the optical fiber is nonlinear and dispersive. An example of the effect of nonlinear distortions is shown in the received constellation in Fig. 4, where the transmission of 16-QAM data is simulated over a particular optical link. The simulation and transmission setup are detailed later in Sec. III. We observe here that because of nonlinear effects, the distortion of each constellation cluster has visually different statistics.

On the one hand, the classification-based equalizer could potentially outperform a regression-based equalizer (in a proper scenario) as it can adapt itself to any channel statistics, so, compared to the MSE regression, the classifier's functioning does not depend on how close the resulting distribution is to the Gaussian one. In particular, $I_{\Theta^*_{\text{cl}}}(X_n; \hat{X}_n)$ can approach the true $I(X;Y)$ provided that the classification equalizer $f_{\text{cl}}(\cdot;\Theta)$ has large enough "learning complexity/capacity" (e.g. number of parameters and flexibility of $f_{\text{cl}}(\cdot;\Theta)$ are sufficient to represent the given phenomenon). However, a classification-based equalizer can also have drawbacks. For instance, one potential drawback is that the CEL function is independent of the spatial proximity of constellation points involved in the decision (classification) process. In other words, the CEL penalizes the misclassification between two classes with the same "cost" value with disregard to the "type" of misclassification occurrence, and this can degrade the NN learning process. This issue is discussed in more detail in [33].

Yet another disadvantage of CEL minimization is that according to [32], the CEL surfaces have fewer local minima than the square error-based losses (SEL), to which the MSE loss belongs. However, the CEL landscape is typically prone to sharp local minima, while for the MSE, these are usually the wide local minima. Such sharp local minima produced by the CEL, exhibit stronger gradients in low training error regions than the MSE does, which causes the infamous overfitting in the systems trained with the CEL, resulting in MSE having a better generalization property in almost all the cases tested in [32].[3] In the next section, we compare the performance and the training pitfalls related to the depth of minima for the classification and regression in the framework of soft-demapping in optical communications.

## *B. Deep Neural Network Design*

Before moving on to the results section, we discuss how we make the comparison of regression and classification as fair as possible. First, we use two types of equalizers: the first is based on the feed-forward multi-layer perceptron (MLP) with three hidden layers, while the second one is the recurrent structure, consisting of one layer of bidirectional Long short-term memory (biLSTM). The comparison of these equalizers' complexity and functioning is given in [15]. These two cases are taken to demonstrate that our outcomes are true for different NN architectures, but we note that the biLSTM layer has demonstrated better performance in previous studies [15]. The only differences between using each architecture for regression or classification tasks occur in i) the structure of the output layer, as shown in Fig 5, and ii) in the loss function type used for each task. In the case of regression, the output layer has two linear neurons referring to the real and imaginary parts of the recovered symbol, and the loss function used is the MSE. For the multi-class classification, the number of neurons in the output layer is determined by the modulation format cardinality (MF or #QAM), and the NN structure ends with the softmax layer, while the loss function is the categorical CEL. We point out once again that besides these two differences, the regression and classifier models that we compare, share the same number of inputs, hidden layers, neurons in each layer, and hyperparameter values; the training/test datasets are also the same. As for the memory sizes, for both types of prediction modeling, we used the same memory length: $M = 51$ for the SSMF case, and $M = 41$ for the TWC case. The values of the NN parameters used in this study are given in Table. I. Furthermore, in terms of training and testing datasets, both NN models were trained and tested with the same datasets, where for training the original training dataset had $2^{20}$ input points, and at every epoch, we picked $2^{18}$ random input points (out of total $2^{20}$) to train the NN

[3] Another drawback of classification-based systems is that it is tailored to each task, i.e. it has to have a specified fixed number of outputs corresponding to the constellation's cardinality. This indicates that the classifier model's operation is specific to the modulation format on which it was trained, the feature that inhibits the practical (say, hardware) classifier implementation. But in the case of regression [47], [48], we do not need to retrain the model at all for it to work on other modulation formats. Because of this, the regression model is far more adaptive than the classification one.

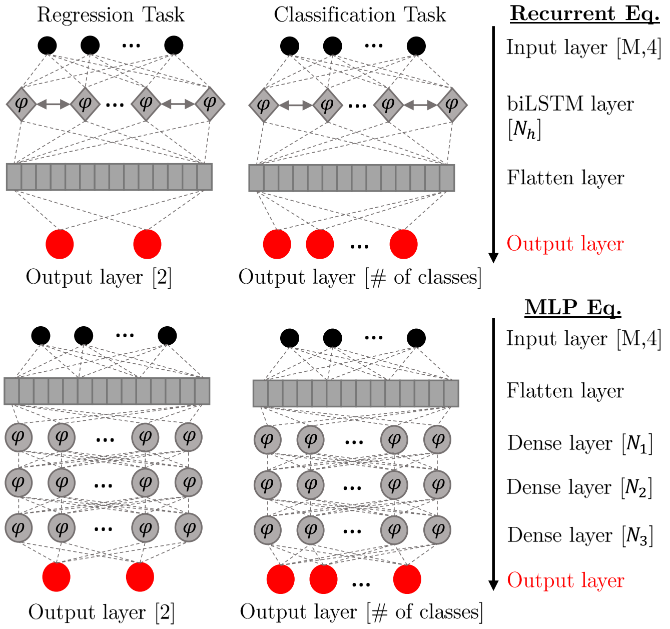


Fig. 5. The schematics of different NN architectures considered in our paper. At the top, we show the regression and classification systems based on the recurrent equalizer with $N_h$ hidden units. At the bottom, we show both tasks implemented with the MLP equalizer having three hidden layers, with $N_1$, $N_2$, and $N_3$ neurons in each consecutive layer, respectively. In all cases, the output is marked with red to highlight the difference in the regression- and multi-class classifier-based approaches. For our case, the activation function $\varphi$ is "tanh". The remarks at the right depict the details of each structure.

TABLE I
THE COMMON REGRESSION AND CLASSIFICATION TASKS NN/TRAINING/TESTING (HYPER)PARAMETERS USED IN OUR STUDY

| **Equalizer** | **Mini-Batch** | $N_h$ | $N_1$ / $N_2$ / $N_3$ | **Training / Testing Dataset size** |
|---|---|---|---|---|
| Recurrent | 4331 | 226 | - | $10^{18}$ / $10^{18}$ |
| MLP | 4331 | - | 481 / 31 / 263 | $10^{18}$ / $10^{18}$ |

efficiently.[4] For testing, we used a never-seen dataset with $2^{18}$ input points.

As we are dealing with different loss functions, we must consider that the learning rate and the number of epochs required, may differ for the regression compared to the classification. To address this potential issue, we optimize the learning rate using values in the range [$10^{-3}$, $5 \cdot 10^{-4}$, $10^{-4}$, $5 \cdot 10^{-5}$], and use the early stop to obtain the architecture that performed best during the training process. In general, the early stop was used if no improvement was seen after 150 epochs out of the total 5000 training epochs.

Additionally, in this work, all NN-based equalizers were implemented in TensorFlow (2.2.0) GPU backend and Keras (2.3.1). Moreover, the Flatten layer, being a part of a Keras API, is used to interconnect the multidimensional (3D) layers with linear (2D - Dense) layers inside the NNs architectures. The codes and dataset samples can be downloaded in [46].

## III. Simulating Signal Propagation in Coherent Optical Transmission Systems

To illustrate the effects addressed in our work, we numerically simulated the dual-polarization (DP) transmission of a single-channel signal propagation at a 34.4 GBd rate. First, a bit sequence was generated using the Mersenne twister generator [52], which has a periodicity equal to $2^{19937} - 1$. Then, the signal is pre-shaped with a root-raised cosine (RRC) filter with 0.1 roll-off at an upsampling rate of 8 samples per symbol. In addition, the signal could have three possible modulation formats: 16 / 32 / 64-QAM. To cover different physical scenarios, we consider the following two test cases: (i) the transmission over the optical link consisting of 9×50 km true-wave classic (TWC) spans; and (ii) the transmission over 5× 100 km of standard single-mode fiber (SSMF) spans. The optical signal propagation along the fiber was simulated by solving the Manakov equation via split-step Fourier method [53] with the resolution of 1 km per step.[5] The parameters of the TWC fiber are: the attenuation parameter $\alpha = 0.23$ dB/km, the dispersion coefficient $D = 2.8$ ps/(nm·km), and the effective nonlinearity coefficient $\gamma = 2.5$ $(\text{W·km})^{-1}$. The SSMF parameters are: $\alpha = 0.2$ dB/km, $D = 17$ ps/(nm·km), and $\gamma = 1.2$ $(\text{W·km})^{-1}$. The purpose of testing two different fibers is to see if the MSE-based regression task works better in the SSMF transmission because, due to the higher dispersion and lower nonlinearity, in that case, the constellation point distributions should be closer to Gaussian; for the TWC we have 6 times lower dispersion and 2 times higher nonlinearity, such that the non-Gaussian constellations should become more pronounced.

In our model, every span is followed by an optical amplifier with the noise figure NF = 4.5 dB, which fully compensates for the fiber losses and adds the ASE noise. At the receiver, a standard Rx-DSP was used. It includes the full electronic chromatic dispersion compensation (CDC) using a frequency-domain equalizer, the application of a matched filter, and downsampling to the symbol rate. Finally, the received symbols were normalized (by phase and amplitude) to the transmitted ones, i.e. multiplied by a constant $\mathcal{K}_{\text{DSP}} \in \mathbb{C}$ minimizing the mean squared error between the transmitted $X_n$ and the received $Y_n$ signals:

$$\mathcal{K}_{\text{DSP}} = \min_{\mathcal{K}} \|\mathcal{K} \cdot Y_n - X_n\|. \tag{9}$$

After Rx-DSP, the output symbols were processed by an NN-based equalizer for further signal enhancement. Fig. 6 shows all the blocks involved in the transmission simulations, where we highlight regression/classification-based NN equalizers/soft-demapping with red boxes. In addition to the AIR, another performance metric used in this paper is the Q-factor calculated directly from the BER values after the hard decision as:

$$Q = 20 \log_{10} \left[ \sqrt{2}\, \text{erfc}^{-1}(2 * \text{BER}) \right], \tag{10}$$

[4] We observed that this training methodology improves the generalization of the training and this was also observed by similar works [49], [50], [51].

[5] The ultra-fine step resolution guarantees that we truly model the optical channel properties captured by the NN and do not address some by-side simulation effects.

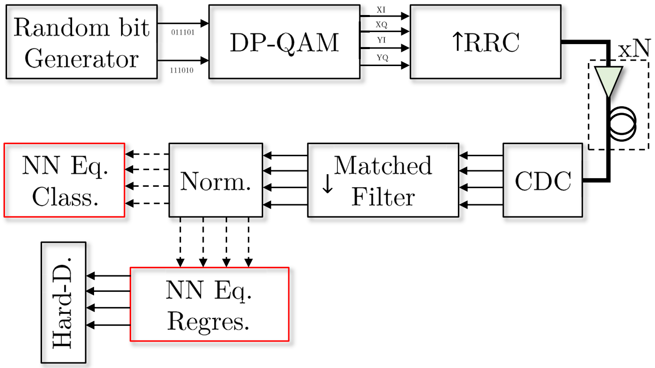


Fig. 6. The schematic of the setup used in our simulations. The two available equalization/soft-demapping types (multi class classification and regression) are inserted at the receiver side after the matching filter and the DSP blocks: CDC and Phase/Amplitude Normalization. The equalizers are highlighted with a red box.

where $\mathrm{erfc}^{-1}$ is the inverse complementary error function. Note that the hard-decision block is optional: it is used for the Q-factor computation, but redundant when we deal with the AIR.

The NN input mini-batch shape, for both regression and classification tasks, can be defined by three dimensions [15]: $(B, M, 4)$, where $B$ is the mini-batch size, $M$ is the memory size defined through the number of neighbours $N$ as $M = 2N + 1$, and $4$ is the number of features for each symbol, referring to the real and imaginary parts of two polarization components. For the regression, the output target is to recover the real and imaginary parts of the $k$-th symbol in one of the polarization, so the shape of the NN output batch can be expressed as $(B, 2)$. In the case of multi-class classification, the output will provide the vector probability of a received symbol to belong to a certain class, and so the output batch shape is equal to $(B, \mathrm{MF})$..[6] Here, we would like to draw your attention to the fact that, since the NN model comes after the traditional Rx-DSP chain, it will only take care of some residual memory and its coupling with the nonlinearity, not the full memory. Finally, we note that different random seeds were used to produce training and testing datasets, ensuring their independence and avoiding overestimation, with the cross-correlation not exceeding $0.02$.

## IV. Comparison of Soft-Demapping Based on Regression or Classification

### A. Performance Comparison

To test the importance of differentiating the output label misclassification, which takes place in the regression as opposed to the multi-class classification, we numerically simulated different modulation formats (16-QAM, 32-QAM, and 64-QAM) transmissions of a single-channel (SC) DP 34.4 GBd signal with RRC 0.1 roll-off pulse over a system consisting of $5 \times 100$ km SSMF spans at 6 and 10 dBm launch power.[7] In this first test, since only the modulation format changes, but we do not have any difference in nonlinearity strength, the goal is to show that when we increase the modulation format, the MC classifier's performance degrades because the categorical cross-entropy loss loses the information related to the miss-classification of different label types and values every misclassification occurrence equally.

The comparison of regression-based and multi-class classification-based systems' performance for different modulation format orders is depicted in Fig. 7. From the results of Fig. 7 (a) and (c), one can see the impact on the MC Classifier's performance when increasing the number of classes in the problem (i.e. increasing the modulation format order), in terms of the Q-factor for 6 and 10 dBm, respectively.

For the biLSTM architecture, the percentage of how higher the Q factor is after regression equalization compared to the multi-class classification one was roughly 8%, 31%, and 31% for the 16-QAM, 32-QAM, and 64-QAM scenarios, respectively, for the 6 dBm test case in Fig. 7 (a), and 22%, 29%, and 50% for the 16-QAM, 32-QAM, and 64-QAM scenarios, respectively, for the 10 dBm test case in Fig. 7 (c). For the MLP architecture, as compared to the multi-class classification output in 16-QAM, 32-QAM, and 64-QAM scenarios, the regression equalization always delivered better results, yielding 14%, 18%, and 15% Q-factor improvement, respectively for the 6 dBm test case, Fig. 7 (a), and 7%, 11%, and 19% Q-factor improvement, respectively, for the 10 dBm test case, Fig. 7 (c). When using the biLSTM layer, we can observe a greater difference between the regression and classification for different modulation formats, because the biLSTM layer, on average, performs much better than just MLP layers [15], [54]. So, in the biLSTM case, we can see better how much the classification loss function gets degraded by ignoring the difference between distinct miss-classification occurrences.

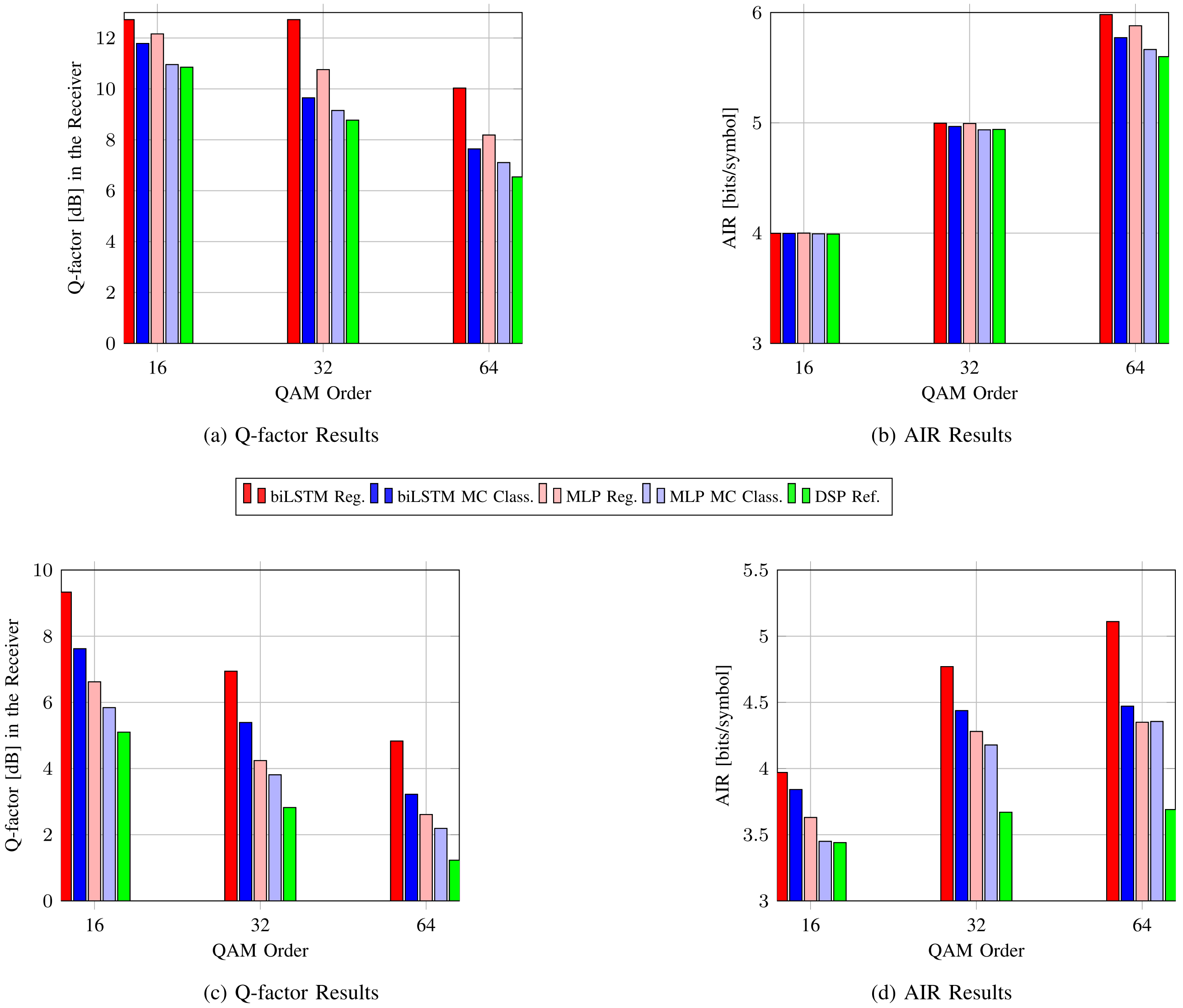
When evaluating the performance in terms of AIR, almost the same behavior was observed: the results are depicted in Fig. 7 (b) and (d) for 6 and 10 dBm, respectively. In the case of the biLSTM architecture, the difference between the AIR obtained by regression and the multiclass classifier for 16-, 32-, and 64-QAM was approximately 0.0008, 0.029, and 0.21, for 6 dBm; and 0.13, 0.33, and 0.63, for 10 dBm. Again, by increasing the order of the modulation format, the regression achieved better results than the multi-class classifiers.[8]

In the case of 6 dBm with the MLP architecture, the same trend appeared again: the AIR difference was larger when the modulation order increased. The difference between AIR for regression vs. multi-class classification was 0.006, 0.057, and 0.21, for modulation formats 16-, 32-, and 64-QAM, respectively. However, for the 10dBm with MLP 64 QAM case (Fig. 7 (d)), the regression and multiclass classifier AIR bars

[6] For the case of multi-label classification as represented in Fig.5, we have appended one extra layer with $\log_2(\mathrm{MF})$ neurons to represent the bit probabilities. So, for this task, the output batch shape is equal to $(B, \log_2(\mathrm{MF}))$

[7] Those transmission powers were picked since they provided BER different than zero in all three modulation formats and they are in the nonlinear regime.

[8] This trend was also observed in [24] in which they showed that the performance of CEL (classifiers) is worse when increasing the order of the QAM modulation format. In [24], they concluded that more accurate gradients are needed for the case of the higher-order constellation, and with the sigmoid function and noisy gradient estimation, the BCEL failed to optimize the high-dimensional parameters.



(a) Q-factor Results

(b) AIR Results

(c) Q-factor Results

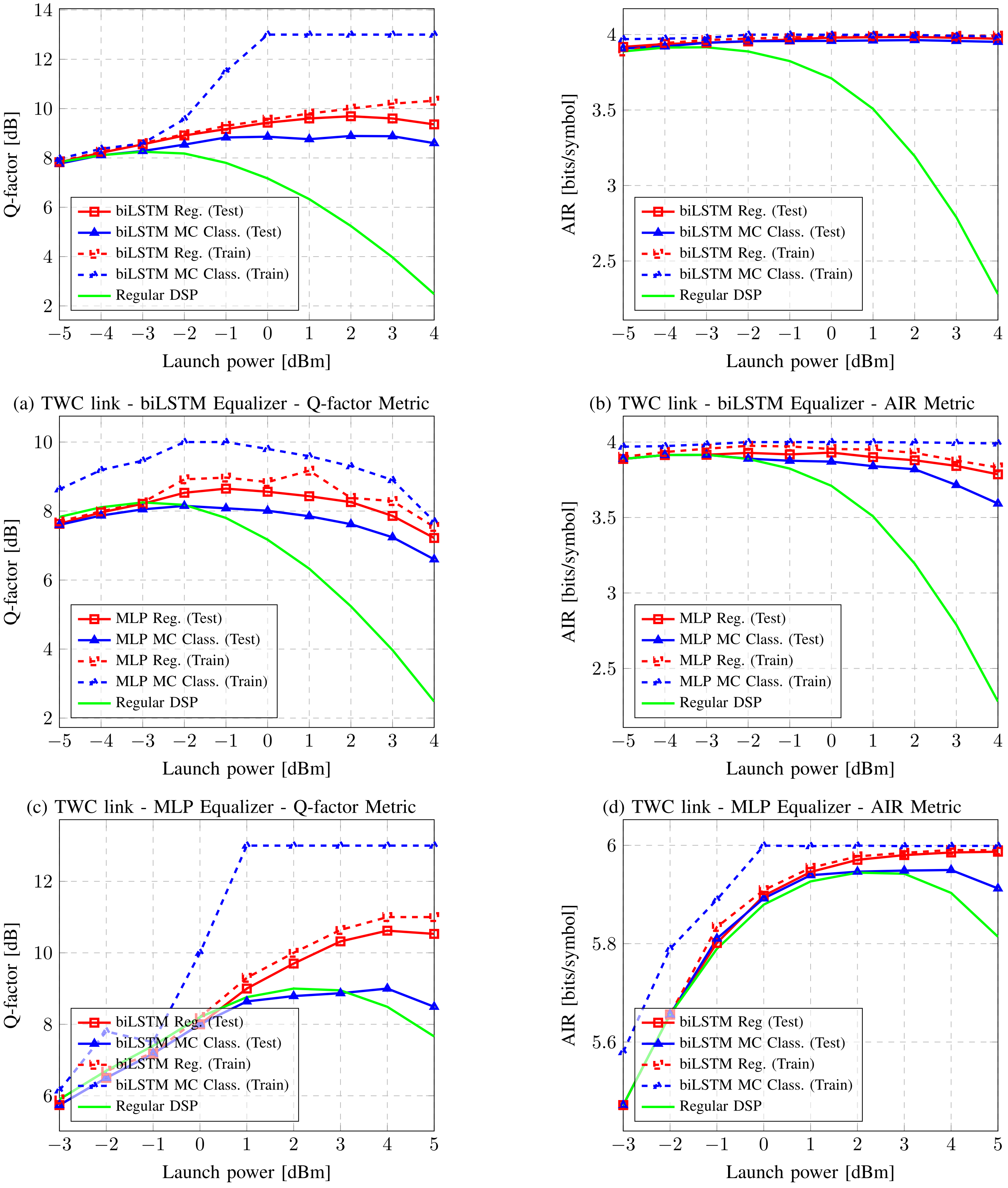
(d) AIR Results

Fig. 7. Performance study for regression equalizer (Reg.) and multi-class classifier (MC Class.) in different modulation format [16-QAM, 32-QAM and 64-QAM] transmission at (a,b) 6 dBm and (c,d) 10 dBm SC-DP, 5 × 100km SSMF fiber and 34.4G Bd.

are nearly identical. We believe that this can be interpreted as a learning limitation of the MLP architecture, and given that the true regression-based AIR value is underestimated, we can confidently state that the overall tendency discovered for all cases studied remains valid.

Now we turn to the issue of overfitting in the multi-class classification equalizer, addressing two cases: i) the case of a low dispersion fiber (SC-DP 16QAM 34.4 GBd over 9 × 50km TWC fiber), and ii) the case of a conventional SSMF fiber (SC-DP 64QAM 34.4 GBd over $5 \times 100$km SSMF fiber). To reveal the overfitting, for both the biLSTM and MLP equalizers, we present the Q-factor/AIR values for the training and test (validation) datasets. The difference between the values obtained in training and testing is the qualitative measure of the overfitting strength: the larger the difference, the stronger the overfitting. Therefore, the comparison of multi-class classification and regression training and testing results will reveal which approach generalizes better. Fig. 8 shows the results of our analysis where the solid green line is the Q-factor/AIR after only linear equalization (regular DSP), the solid blue and red lines indicate the Q-factor/AIR of the multi-class classification and regression models evaluated with the testing dataset, respectively, and the dashed blue and red lines depict the Q-factor/AIR of the multi-class classification and regression models evaluated with the training dataset.

When we use the CEL in our equalizers, we see the same trend of higher overfitting levels as was observed in [32]. As can be seen in all four panels, the Q-factor curves of training and testing for the multi-class classifiers show a significant difference (since this metric gives the logarithmic measure), suggesting the presence of noticeable overfitting in the classification model. Then, we can see that, in comparison to the multi-class classifier's result, the training and testing output curves, when using the regression, behave almost identically. It means that the regression model using MSE generalizes much better for all of our test cases



(a) TWC link - biLSTM Equalizer - Q-factor Metric

(b) TWC link - biLSTM Equalizer - AIR Metric

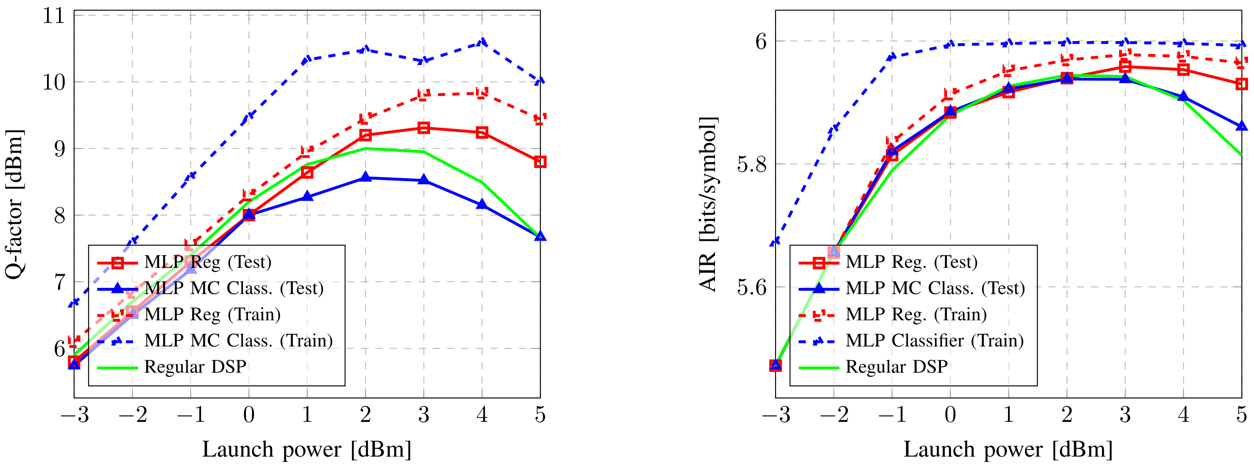
(c) TWC link - MLP Equalizer - Q-factor Metric

(d) TWC link - MLP Equalizer - AIR Metric

(e) SSMF link - biLSTM Equalizer - Q-factor Metric

(f) SSMF link - biLSTM Equalizer - AIR Metric

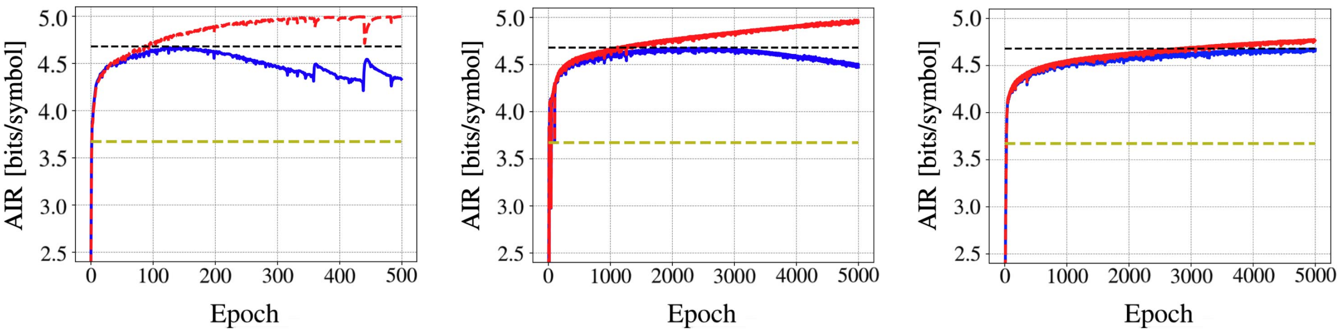
Fig. 8. Generalization study for regression equalizer (Reg.) and Multi-Class Classifier (MC Class.) showing the impact of overfitting in the NN performance and training process on the following scenarios: (a,b) biLSTM analyses and (c,d) MLP analyses for SC-DP-16-QAM, 9 × 50km TWC fiber and 34.4GBd; (e,f) biLSTM analyses and (g,h) MLP analyses for SC-DP-64-QAM, 5 × 100km SSMF fiber and 34.4GBd.



(g) SSMF link - MLP Equalizer - Q-factor Metric

(h) SSMF link - MLP Equalizer - AIR Metric

Fig. 8. *(Continued.)* Generalization study for regression equalizer (Reg.) and Multi-Class Classifier (MC Class.) showing the impact of overfitting in the NN performance and training process on the following scenarios: (a,b) biLSTM analyses and (c,d) MLP analyses for SC-DP-16-QAM, 9 × 50km TWC fiber and 34.4GBd; (e,f) biLSTM analyses and (g,h) MLP analyses for SC-DP-64-QAM, 5 × 100km SSMF fiber and 34.4GBd.



(a)

(b)

(c)

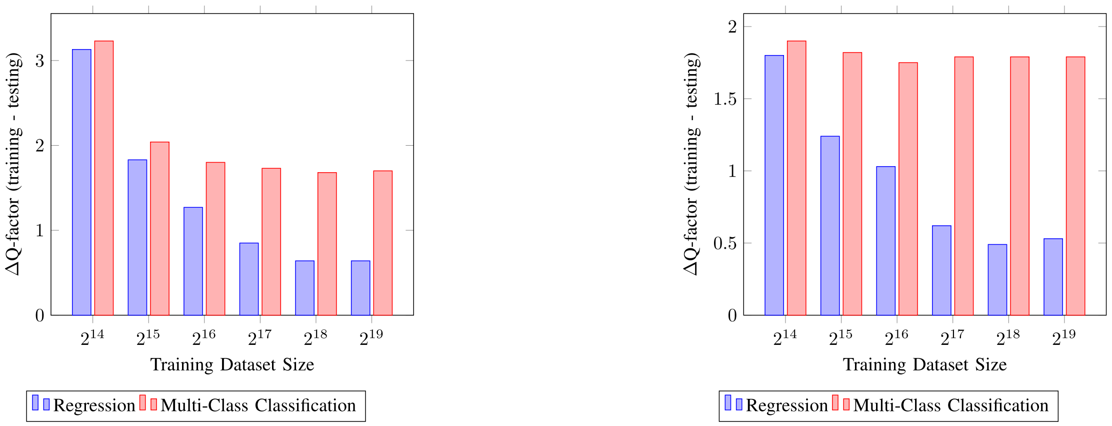
Fig. 9. The performance (AIR) versus training epochs for the 32-QAM SSMF 10 dBm case using the biLSTM Multi-Class Classifier with learning rate equal to (a) $10^{-3}$, (b) $10^{-4}$, and (c) $5 \times 10^{-5}$. The red solid curve is the training performance, the blue solid curve is the test performance, the green dashed line is the reference performance when only the linear equalization is used, and the black dashed line is the reference for the maximum AIR achieved by the testing dataset over the training epochs.

(two different NN equalizers and two different transmission setups), which, again, complies with the conclusions reached in [32]. Furthermore, we were able to see that, by using regression equalizers, the Q-factor level after equalization was still higher than with the classification, due to the better generalization of the regression NN models. When we look at the AIR values, we see that the multi-class classifier's training performance was overfitted, yielding virtually the maximum AIR attainable for each scenario, but in the case of regression, the training and testing curves followed the same trend, indicating a better generalization of the problem. Also, we notice that in the 64-QAM 5 × 100 SSMF transmission scenario, we have an even more reduced classification performance, with no increase in Q-factor observed for both biLSTM and MLP equalizers.

Finally, we would like to note that when we lowered the learning rate, we saw a reduction in overfitting in the classification task. Fig. 9 shows the example case of 32-QAM at 10 dBm using the SSMF link, where training and testing AIR curves for the biLSTM equalizer are shown for the three learning rates: $10^{-3}$, $10^{-4}$, and $5 \times 10^{-5}$. As can be seen, for $10^{-3}$ the overfitting is much more intense than when using lower learning rates. However, even with such low learning rates as $5 \times 10^{-5}$, the performance after 5000 epochs of training, was not better than that with either the regression or compared to the best case with the classification with $10^{-3}$ learning rate. Also, the overfitting could still be seen, as the training AIR level grew faster than the testing level. The maximum AIR measured with the test dataset is shown by the black dashed line in Fig. 9. It is evident from this figure that lowering the learning rate did not result in any dramatic improvement in AIR for our test case. Next, we begin the investigation of the possible causes of the training challenges for the multi-class and multi-label classifiers using the learning rate equal to $10^{-4}$.

### *B. Training Pitfalls and Overfitting Investigation*

In this final subsection, we outline the NN training process employed in this paper, taking into account not only the problem of overfitting but also the possible reasons why the classifiers struggled during learning and did not generalize



(a) TWC link - MLP equalizer

(b) SSMF link - MLP equalizer

Fig. 10. Difference between training and testing performance over different training dataset sizes to evaluate the overfitting behavior. The MLP was applied for 3dBm 34.4GBd SC-DP, 9 × 50km TWC and 5 × 100km SSMF fiber links.

well to the testing dataset. Once again, we emphasize that the CEL is the ideal loss function from the information theory viewpoint and that all of the performance degradation impacts we saw are directly attributable to the classifiers' gradient descent learning.

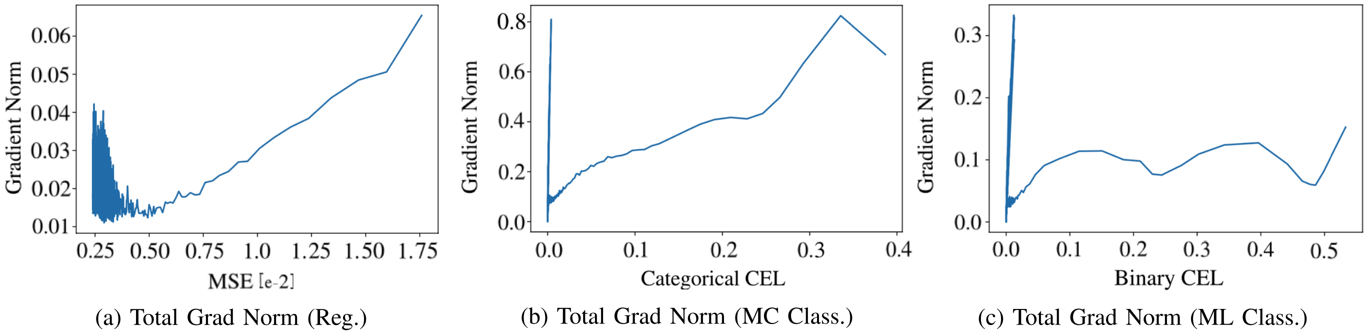
First, it is important to note that one possible criticism could be that the representation capability of the NN part may turn out to be insufficient to perform the demapping in the case of the classifier, while the regression could emerge as a "simpler task", such that the NN part is sufficient and does its job better in the latter case. Even though we do not discard that it might, potentially, play some role in specific cases (though it does not do in this current study), in this paper, we show that some fundamental machine learning-related issues underlie the worsening in the performance of classifiers as compared to the regression-based predictive modelling. Therefore, we claim that the NN‘s capacity in the classification case is not the main problem, but the overfitting and gradient-related issues are the true cause. According to the literature [see, e.g., [26], Chapter 5.2], when an NN model is overfitting, our adding an extra complexity/capacity (say, in the form of additional hidden layers or neurons) would only worsen the overfitting effect. As a form of regularization, it is commonly suggested to do quite the opposite: to reduce the NN size for the sake of mitigating the strong overfitting effects. We have tested increasing the number of neurons in the classifier and observed a similar behaviour (for the same number of epochs) as we had had for the initial number of layers/weights, and Q-factor stayed around the "original" value. Additionally, we notice that even in the original case presented in Fig. 5, the complexity of the regression-based equalizer and classifier is, in fact, not the same, despite their sharing the same NN part. In the regression, we have only two outputs, and in the MC classification, we have QAM-order-number neurons plus the softmax activation function that is responsible for the decision process, making the classification more complex than the regression in terms of multiplications number.

Having understood that the NN's capacity is not the problem's source, we now present the new results to show the true reason why the classifiers have high overfitting and demonstrate poorer performance in our study. The first and common hypothesis of such strong overfitting is that it occurs due to the lack of data (insufficient diversity) used in the training. To show that this is not the root cause of our problems, we have trained the NN classifier and regression equalizers, gradually enlarging the training data size from $2^{14}$ to $2^{19}$ symbols.[9] In Fig.10, we show the difference in the Q factor between the training and testing runs, depending on the size of the dataset. As it is clear from this result, the strong overfitting was observed for the regression when training the system with a lower amount of data, and by increasing the dataset, the difference decreased to approximately 0.5 dB in both SSMF and TWC cases. The latter value can be reckoned as "an acceptable margin" to allow the system's generalization: training and testing give approximately the same result. However, when dealing with the classification, we see that increasing the dataset size is not enough: the difference between the training and testing results was still significant. Even though both regression and classification were trained with the same datasets, we can see that classification overfitting is much more pronounced, especially for the SSMF case. Therefore, the important question raised in our work is: What can be the cause of this overfitting inclination in the classification case?

Since, as we saw, the lack of data and the NN complexity/capacity, are not the main inducement for the classification system overfitting,[10] we believe that the reason for such a behaviour is different. For the classification, the poor performance comes from the fact that for such high levels of accuracy as we have in the communications, the CEL produces really sharp local landscape minima that will cause the infamous problem of vanishing back-propagating gradients, making the classifier effectively cease learning. The fact of the

[9] For this particular study, the data have not been randomly shuffled, and we use the same dataset size on all epochs.

[10] Note that the overfitting problem can be rectified if a very large amount of training data is used or if we apply regularization procedures. However, the above measures do not address the gradient-related issues that are the genuine source of learning degradation for classifiers discussed in this section.



(a) Total Grad Norm (Reg.) (b) Total Grad Norm (MC Class.) (c) Total Grad Norm (ML Class.)

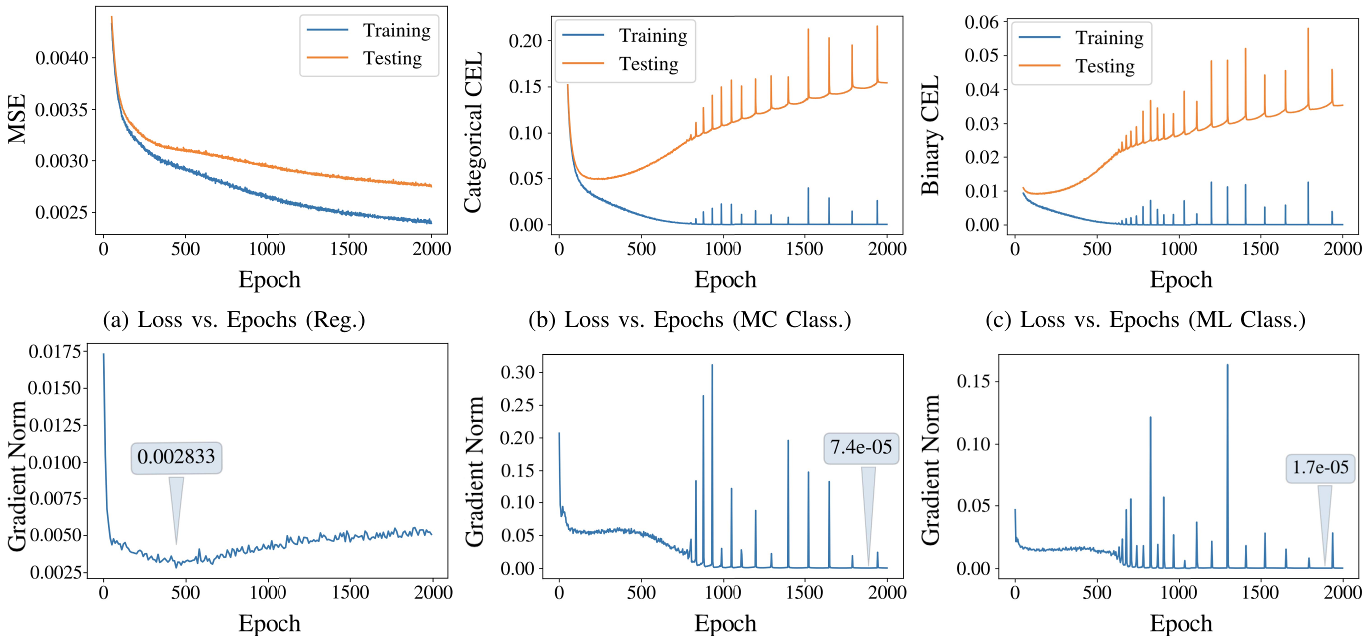
Fig. 11. Gradient norm versus Training loss for the Regression model (a), Multi-Class Classifier (b), and Multi-Label Classifier (c) to investigate the gradient close to the minimum loss.

sharp minima appearance in the CEL landscape can be shown by following the same method as in [32]. To demonstrate this, we plot the gradient norm of the NN structure over the training loss value to evaluate whether a particular local minimum is sharp or wide. Our results are summarized in Fig. 11 (a), (b), and (c), for the regression, multi-class, and multi-label classification, respectively. Thus, in these figures, we recognize the same pattern as observed in [32, Fig. 15]: the CEL revealed multiple points of high gradient close to the loss minima (in our case the gradient norm reaches 0.8, and in [32] it reached 0.5), and it is exactly the very indication of a sharp minimum in the loss landscape. On the other hand, the MSE was much less prone to overfitting due to weaker gradients close to the minima (in our case, those ranged from 0.01 to 0.04, in [32] it reached 0.05). Therefore, our analysis demonstrates the sheer difference in the classification and regression cases: we have sharp local minima in the CEL landscape and do not have those for the MSE loss.

In addition to this analysis, a practical way to demonstrate that local minima cause the CEL learning problem is to verify that the gradient norm shrinks to some "insignificant" values along with the training, as described in [26, Chapter 8.2.2]. In Fig. 12, we present the training and testing loss value over the epochs and the gradient norm of the first layer of each model as a function of the epoch number: the regression (MSE loss) in panels (a), (d), multi-class – in (b), (e), multi-label –in panes (c), (f). In the lower row, we explicitly highlight the minimal gradient value over the epochs. In this gradient study, we consider only the MLP architecture applied to the SSMF fiber case. From this figure, we can clearly see that the gradient, when using the CEL, vanishes, whereas the same did not happen in the MSE loss case. When using the CCEL, the minimum gradient in the first layer reached $7.4e-5$, and when using the BCEL, it dropped further to $1.7e-5$; but, when using the MSE, the minimal gradient was around $0.003$. Here, we emphasize that the gradient value of $10^{-5}$ order is a clear indication of a vanishing gradient problem. Now, looking at the total gradient norm over the entire model (the absolute value of the sum of gradients for all MLP layers), we checked two cases: i) the gradient value at the epoch, where the testing loss value was lowest; ii) the gradient value after 2000 epochs. For the MSE case, the gradient norm in i) was 0.017, and in ii) it was 0.010; in the case of CCEL, the gradient norm in i) was 0.083, and in ii) it was 1.17e-04. Eventually, in the case of BCEL, the gradient norm in i) was 0.024, and in ii) it was 3.12e-05. With this result, we can see that the "gradient shrinking to insignificant values along the training" behaviour mentioned in [26], for both the CCEL and BCEL cases. Also, we point out that the gradient vanishing problem in the softmax with CCEL was observed in [55], and the same for the sigmoid with the BCEL, was reported in [24]. We claim that the same phenomenon happens in communications due to the high accuracy (low error) levels in input datasets that we have to work with, stipulating the difference in the performance of classification and regression-based predictive modelling that we report here.

Additionally, for this same transmission case, Fig. 12, we could see that the original MSE before the NN, was 0.0066, and the CCEL before the NN, assuming the Gaussian distributions of constellation points, was 0.057. After the NN, the MSE of the regression was reduced to 0.0027 ($59\%$ drop in loss), and the CCEL of the multi-class classifier fell down to 0.049 ($15\%$ drop in loss). This fact, together with the gradient analyses, shows that for such high accuracy systems, where at the beginning of the training the loss function value is already very small, the MSE can still avoid vanishing of gradients, while the CEL falls in really sharp local minima and stops learning due to the vanishing gradient problem. Furthermore, we tried to address the overfitting issue with traditional methods (increasing the training dataset size, adding a dropout layer, batch normalization layer, and introducing a regularizer). However, even though the overfitting was partially mitigated, the Q-factor on a testing dataset was still almost the same because the vanishing gradients issue had not been solved with those measures.

Finally, we would like to show the Q-factor performance of the multi-label classification using the BCEL. As we could see in Figs. 11, 12 the BCEL plus sigmoid suffered from the same problems as the CCEL plus a softmax layer, and this translated into the degradation of performance as well. In terms of performance, when the MLP with a learning rate of 0.0001, was applied to the 64QAM, 3dBm, SC-DP, 5 × 100 km 34.4 GBd SSMF link, the maximal Q-factor achieved by the regression was 9.6 dB, for the multi-class classifier it was 8.8dB, and for the multi-label classifier it was just 8.7 dB, showing that the change in the loss function



(a) Loss vs. Epochs (Reg.)

(b) Loss vs. Epochs (MC Class.)

(c) Loss vs. Epochs (ML Class.)

(d) Grad Norm of $1^o$ layer (Reg.)

(e) Grad Norm of $1^o$ layer (MC Class.)

(f) Grad Norm of $1^o$ layer (ML Class.)

Fig. 12. Overfitting and Gradient Vanishing investigation for the Regression model (a/d), Multi-Class Classifier (MC Class.) (b/e), and Multi-Label Classifier (MB Class.) (c/f).

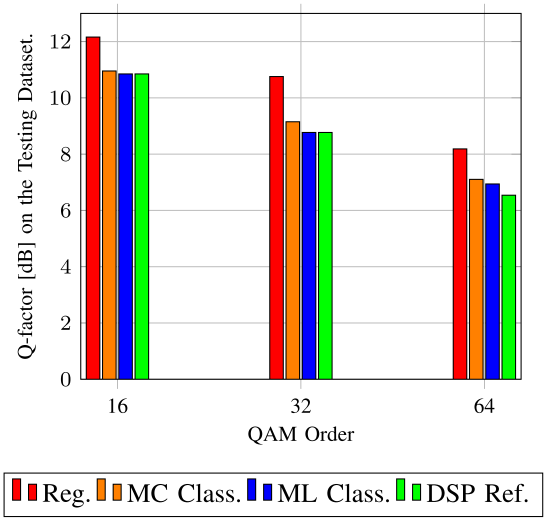


Fig. 13. Performance evaluation on the testing datasets for three different modulation formats (16-,32-, and 64-QAM). The MLP was applied for 6dBm SC-DP, 5 × 100km 34.4GBd SSMF fiber links.

and the addition of one extra layer did not help improve BER. Also, we appended the results in Fig 7 a) with those referring to the multi-label classification: it is depicted in Fig. 13. There, we can see a trend similar to that reported in [24]: the regression equalization always delivered better results, yielding 14%, 18%, and 15% Q-factor improvement for 16-QAM, 32-QAM, and 64-QAM, respectively, compared to the multi-class classification output. Turning to the BCEL case (multi-label classifier), this difference becomes even greater: the regression equalization improved the Q-factor by 15%, 23%, and 18%, for the 16-QAM, 32-QAM, and 64-QAM scenarios, respectively.

We conclude this section with a comparison to a typical computer vision challenge and our view on the next steps in studying the loss function landscape occurring in communications-related applications. In computer vision, it was discovered that when an image is processed, the bulk of the pixies describes a graphic background, and only a few pixels pertain to the genuine objects in the image. This resulted in inefficient training because most parts of the image correspond to "an easy prediction" (which means that they can be easily labeled as background by the detector) and therefore offer little relevant learning. Although they individually provide tiny contributions to the loss value, when we combine those contributions, they can overwhelm the loss and computed gradients, resulting in a degraded model's prediction performance, since easy predictions (detections with high probabilities, or, in our context, the correct classifications following a simple hard decision) account for a large share of inputs. To address this issue, in [41] Facebook A. I. developed a new modified approach named focal loss (FL), by adding a weighting factor to the cross-entropy loss. The FL gives a higher weight to cases that are hardly misclassified: in communications, it would correspond to the cases that got misclassified after the hard-decision (HD) process. We believe that the difficulty of the "dataset imbalance" (meaning that just a small fraction of the dataset corresponds to the wrong HD predictions) exists in the high-accuracy communication-related equalization/demapping problem. As a toy example, consider a system where an initial SER after the HD, is equal to $10^{-3}$. Training the NN-classifier, in this case, will mean that a 99.9% fraction of the training dataset corresponds to "an easy prediction", and the remaining 0.1% will be the "hard prediction" members. Therefore, to improve the performance of classifiers, we expect that a similar focal loss function, as in computer vision, must be created for the communications application.

## V. Conclusion

In this work, we compared the performance and training peculiarities of the regression and classification predictive models, addressing the NN-based soft-demapping in coherent optical communication. We considered several transmission scenarios, including three different modulation formats, on two different optical link test benches with different nonlinear and dispersion responses. The applied NN models were based on two different architectures: the feedforward and recurrent NNs. For the regression equalizer and the multi-class classifier, the model had the same structure, except for the configuration of the last layer conditioned by the particular predictive modeling tasks. In most of the scenarios studied in this work, the soft-demapping based on regression outperforms the one based on classification providing higher Q-factor and achievable information rate. We have further observed that the soft-demapping based on cross-entropy learning required was more prone to overfitting than the regression-based counterparts. This observation regarding overfitting is in line with findings from [32] showing the performance advantage of regression models over classification models in a different context, due to the better generalization capability of the former.

We should emphasize that both the regression and classification tasks have certain limitations. The regression loss function (the MSE) is a special case of the classification loss function (the CEL), in which the stochastic component of the output variables is assumed to be signal-independent and normally distributed. Therefore, the MSE does not take into account the signal-dependent stochastic contribution, which is, obviously, present in the true nonlinear optical channel.

Nevertheless, we underline that from the machine learning methods' application perspective, the classification loss function (the CEL) landscape typically involves very sharp local minima, which can cause the NN model to generalize much worse than the regression model with the loss based on the Euclidean distance. Additionally, we show that due to the high accuracy transmissions involved in communications, the CEL tends to vanish the gradients while the MSE can still maintain a staple gradient landscape that makes the NN learn more effectively.

Finally, although we have observed a common performance trend between these two predictive models, a general conclusion cannot be made. For instance, some advanced problem-specific loss functions or data pre-processing/mining may potentially hinder the training problems in the classification case. However, the scope of this work was to evaluate the performance of the typical classifier and regression architectures with their conventional loss function and we leave it to future investigation on how to improve the training process on NN classifiers in the task of soft demapping.

## Acknowledgment

The authors would like to thank the associate editor and unanimous reviewers for insightful discussions and suggestions.

## References

[1] E. Agrell et al., "Roadmap of optical communications," *J. Opt.*, vol. 18, no. 6, May 2016, Art. no. 063002.

[2] P. J. Winzer, D. T. Neilson, and A. R. Chraplyvy, "Fiber-optic transmission and networking: The previous 20 and the next 20 years," *Opt. Exp.*, vol. 26, no. 18, pp. 24190–24239, 2018.

[3] J. C. Cartledge, F. P. Guiomar, F. R. Kschischang, G. Liga, and M. P. Yankov, "Digital signal processing for fiber nonlinearities," *Opt. Exp.*, vol. 25, no. 3, pp. 1916–1936, Feb. 2017.

[4] F. N. Khan, C. Lu, and A. P. T. Lau, "Machine learning methods for optical communication systems," in *Signal Processing in Photonic Communications*. Washington, DC, USA: Optical Society of America, 2017.

[5] D. Zibar, M. Piels, R. Jones, and C. G. Schaeffer, "Machine learning techniques in optical communication," *J. Lightw. Technol.*, vol. 34, no. 6, pp. 1442–1452, Mar. 15, 2016.

[6] F. Musumeci et al., "An overview on application of machine learning techniques in optical networks," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 2, pp. 1383–1408, 2nd Quart., 2019.

[7] M. A. Jarajreh et al., "Artificial neural network nonlinear equalizer for coherent optical OFDM," *IEEE Photon. Technol. Lett.*, vol. 27, no. 4, pp. 387–390, Feb. 15, 2015.

[8] D. Wang et al., "System impairment compensation in coherent optical communications by using a bio-inspired detector based on artificial neural network and genetic algorithm," *Opt. Commun.*, vol. 399, pp. 1–12, Sep. 2017.

[9] S. Zhang et al., "Field and lab experimental demonstration of nonlinear impairment compensation using neural networks," *Nature Commun.*, vol. 10, no. 1, pp. 1–8, 2019.

[10] P. J. Freire et al., "Complex-valued neural network design for mitigation of signal distortions in optical links," *J. Lightw. Technol.*, vol. 39, no. 6, pp. 1696–1705, Dec. 3, 2021.

[11] C. Häger and H. D. Pfister, "Nonlinear interference mitigation via deep neural networks," in *Proc. Opt. Fiber Commun. Conf.*, 2018, pp. 1–3.

[12] B. I. Bitachon, A. Ghazisaeidi, M. Eppenberger, B. Baeurle, M. Ayata, and J. Leuthold, "Deep learning based digital backpropagation demonstrating SNR gain at low complexity in a 1200 km transmission link," *Opt. Exp.*, vol. 28, no. 20, pp. 29318–29334, Sep. 2020.

[13] M. Schaedler, F. Pittala, G. Bocherer, C. Bluemm, M. Kuschnerov, and S. Pachnicke, "Recurrent neural network soft-demapping for nonlinear ISI in 800 Gbit/s DWDM coherent optical transmissions," in *Proc. Eur. Conf. Opt. Commun. (ECOC)*, Dec. 2020, pp. 1–4.

[14] Y. Zhao et al., "Low-complexity fiber nonlinearity impairments compensation enabled by simple recurrent neural network with time memory," *IEEE Access*, vol. 8, pp. 160995–161004, 2020.

[15] P. J. Freire et al., "Performance versus complexity study of neural network equalizers in coherent optical systems," *J. Lightw. Technol.*, vol. 39, no. 19, pp. 6085–6096, Oct. 1, 2021.

[16] T. J. O'Shea and J. Hoydis, "An introduction to deep learning for the physical layer," *IEEE Trans. Cogn. Commun. Netw.*, vol. 3, no. 4, pp. 563–575, Oct. 2017.

[17] B. Karanov, P. Bayvel, and L. Schmalen, "End-to-end learning in optical fiber communications: Concept and transceiver design," in *Proc. Eur. Conf. Opt. Commun. (ECOC)*, Dec. 2020, pp. 1–4.

[18] B. Karanov et al., "End-to-end deep learning of optical fiber communications," *J. Lightw. Technol.*, vol. 36, no. 20, pp. 4843–4855, Oct. 15, 2018.

[19] Z.-R. Zhu, J. Zhang, R.-H. Chen, and H.-Y. Yu, "Autoencoder-based transceiver design for OWC systems in log-normal fading channel," *IEEE Photon. J.*, vol. 11, no. 5, pp. 1–12, Oct. 2019.

[20] K. Gümüş, A. Alvarado, B. Chen, C. Häger, and E. Agrell, "End-to-end learning of geometrical shaping maximizing generalized mutual information," in *Proc. Opt. Fiber Commun. Conf. (OFC)*, 2020, pp. 1–3.

[21] F. A. Aoudia and J. Hoydis, "End-to-end learning for OFDM: From neural receivers to pilotless communication," 2020, *arXiv:2009.05261*.

[22] V. Neskorniuk et al., "End-to-end deep learning of long-haul coherent optical fiber communications via regular perturbation model," in *Proc. ECOC*, 2021, pp. 1–4.

[23] V. Aref and M. Chagnon, "End-to-end learning of joint geometric and probabilistic constellation shaping," in *Proc. Opt. Fiber Commun. Conf. (OFC)*, 2022, pp. 1–3.

[24] Z. Niu, H. Yang, H. Zhao, C. Dai, W. Hu, and L. Yi, "End-to-end deep learning for long-haul fiber transmission using differentiable surrogate channel," *J. Lightw. Technol.*, vol. 40, no. 9, pp. 2807–2822, May 1, 2022.

[25] G. Böcherer. (2021). *Lecture Notes on Machine Learning for Communications*. [Online]. Available: http://georg-boecherer.de/mlcomm

[26] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016. [Online]. Available: http://www.deeplearningbook.org

[27] S. Dörner, S. Cammerer, J. Hoydis, and S. Ten Brink, "Deep learning based communication over the air," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 1, pp. 132–143, Feb. 2017.

[28] S. Deligiannidis, A. Bogris, C. Mesaritakis, and Y. Kopsinis, "Compensation of fiber nonlinearities in digital coherent systems leveraging long short-term memory neural networks," *J. Lightw. Technol.*, vol. 38, no. 21, pp. 5991–5999, Nov. 1, 2020.

[29] M. Schädler, G. Böcherer, and S. Pachnicke, "Soft-demapping for short reach optical communication: A comparison of deep neural networks and Volterra series," *J. Lightw. Technol.*, vol. 39, no. 10, pp. 3095–3105, May 15, 2021.

[30] P. Golik, P. Doetsch, and H. Ney, "Cross-entropy vs. squared error training: A theoretical and experimental comparison," in *Proc. Interspeech*, vol. 13, Aug. 2013, pp. 1756–1760.

[31] K. Nar, O. Ocal, S. S. Sastry, and K. Ramchandran, "Cross-entropy loss and low-rank features have responsibility for adversarial examples," 2019, *arXiv:1901.08360*.

[32] A. S. Bosman, A. Engelbrecht, and M. Helbig, "Visualising basins of attraction for the cross-entropy and the squared error neural network loss functions," *Neurocomputing*, vol. 400, pp. 113–136, Aug. 2020.

[33] S. Lathuilière, P. Mesejo, X. Alameda-Pineda, and R. Horaud, "A comprehensive analysis of deep regression," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 9, pp. 2065–2081, Sep. 2019.

[34] Y. Li, C. Lan, J. Xing, W. Zeng, C. Yuan, and J. Liu, "Online human action detection using joint classification-regression recurrent neural networks," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 203–220.

[35] L. Mukherjee, M. A. K. Sagar, J. N. Ouellette, J. J. Watters, and K. W. Eliceiri, "Joint regression-classification deep learning framework for analyzing fluorescence lifetime images using nadh and fad," *Biomed. Opt. Exp.*, vol. 12, no. 5, pp. 2703–2719, 2021.

[36] M. Liu, J. Zhang, E. Adeli, and D. Shen, "Joint classification and regression via deep multi-task multi-channel learning for Alzheimer's disease diagnosis," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 5, pp. 1195–1206, May 2018.

[37] M. Liu, J. Zhang, E. Adeli, and D. Shen, "Deep multi-task multi-channel learning for joint classification and regression of brain status," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*, 2017, pp. 3–11.

[38] J.-Y. Wu, M. Wu, Z. Chen, X. Li, and R. Yan, "A joint classification-regression method for multi-stage remaining useful life prediction," *J. Manuf. Syst.*, vol. 58, pp. 109–119, Jan. 2021.

[39] J. Chen, L. Cheng, X. Yang, J. Liang, B. Quan, and S. Li, "Joint learning with both classification and regression models for age prediction," *J. Phys., Conf.*, vol. 1168, Feb. 2019, Art. no. 032016.

[40] P. J. Freire, A. Napoli, B. Spinnler, N. Costa, S. K. Turitsyn, and J. E. Prilepsky, "Neural networks-based equalizers for coherent optical transmission: Caveats and pitfalls," *IEEE J. Sel. Topics Quantum Electron.*, vol. 28, no. 4, pp. 1–23, Jul. 2022.

[41] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[42] P. Sadeghi, P. O. Vontobel, and R. Shams, "Optimization of information rate upper and lower bounds for channels with memory," *IEEE Trans. Inf. Theory*, vol. 55, no. 2, pp. 663–688, Feb. 2009.

[43] G. Böcherer, P. Schulte, and F. Steiner, "Probabilistic shaping and forward error correction for fiber-optic communication systems," *J. Lightw. Technol.*, vol. 37, no. 2, pp. 230–244, Jan. 28, 2019.

[44] N. Merhav, G. Kaplan, A. Lapidoth, and S. Shamai, "On information rates for mismatched decoders," *IEEE Trans. Inf. Theory*, vol. 40, no. 6, pp. 1953–1967, Nov. 1994.

[45] F. Diedolo, G. Böcherer, M. Schädler, and S. Calabró, "Nonlinear equalization for optical communications based on entropy-regularized mean square error," in *Proc. Eur. Conf. Opt. Commun. (ECOC)*, 2022, pp. 1–4.

[46] P. J. Freire. (Feb. 2022). *Code for Regression Equalizer, Multi-Class Classifier and Multi-Label Classifier*. [Online]. Available: https://doi.org/10.5281/zenodo.7004204

[47] P. J. Freire et al., "Transfer learning for neural networks-based equalizers in coherent optical systems," *J. Lightw. Technol.*, vol. 39, no. 21, pp. 6733–6745, Nov. 1, 2021.

[48] P. J. Freire, D. Abode, J. E. Prilepsky, and S. K. Turitsyn, "Power and modulation format transfer learning for neural network equalizers in coherent optical transmission systems," in *Proc. OSA Adv. Photon. Congr.*, 2021, Art. no. SpM5C-6.

[49] E. Hoffer, T. Ben-Nun, I. Hubara, N. Giladi, T. Hoefler, and D. Soudry, "Augment your batch: Improving generalization through instance repetition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8129–8138.

[50] S. Kumawat, G. Kanojia, and S. Raman, "ShuffleBlock: Shuffle to regularize deep convolutional neural networks," 2021, *arXiv:2106.09358*.

[51] Z.-L. Ke, H.-Y. Cheng, and C.-L. Yang, "LIRS: Enabling efficient machine learning on NVM-based storage via a lightweight implementation of random shuffling," 2018, *arXiv:1810.04509*.

[52] M. Matsumoto and T. Nishimura, "Mersenne twister: A 623-dimensionally equidistributed uniform pseudo-random number generator," *ACM Trans. Model. Comput. Simul.*, vol. 8, no. 1, pp. 3–30, Jan. 1998.

[53] G. P. Agrawal, *Nonlinear Fiber Optics*, 5th ed. Boston, MA, USA: Academic, 2013.

[54] P. J. Freire et al., "Experimental study of deep neural network equalizers performance in optical links," in *Proc. Opt. Fiber Commun. Conf. Exhib. (OFC)*, 2021, pp. 1–3.

[55] S. Wang, F. Liu, and B. Liu, "Escaping the gradient vanishing: Periodic alternatives of softmax in attention mechanism," *IEEE Access*, vol. 9, pp. 168749–168759, 2021.