IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 6, 2021, accepted August 31, 2021, date of publication September 6, 2021, date of current version September 16, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3110647

Normalization Influence on ANN-Based Models Performance: A New Proposal for Features' Contribution Analysis

IRATXE NIÑO-ADAN[1,2], EVA PORTILLO[2], ITZIAR LANDA-TORRES[3], AND DIANA MANJARRES[1]

[1]Tecnalia Research and Innovation, Basque Research and Technology Alliance (BRTA), 48160 Derio, Spain
[2]Department of Automatic Control and Systems Engineering, Faculty of Engineering, University of the Basque Country (UPV/EHU), 48013 Bilbao, Spain
[3]Petronor Innovación S.L., 48550 Muskiz, Spain

Corresponding author: Iratxe Niño-Adan (iratxe.nino@tecnalia.com)

This work was supported in part by DATA Inc. Fellowship under Grant 48-AF-W1-2019-00002, in part by Tecnalia Research and Innovation Ph.D. Scholarship, in part by the Spanish Centro para el Desarrollo Tecnológico Industrial (CDTI, Ministry of Science and Innovation) through the ''Red Cervera'' Programme (AI4ES Project) under Grant CER-20191029, and in part by the 3KIA Project funded by the ELKARTEK Program of the SPRI-Basque Government under Grant KK-2020/00049.

ABSTRACT Artificial Neural Networks (ANNs) are weighted directed graphs of interconnected neurons widely employed to model complex problems. However, the selection of the optimal ANN architecture and its training parameters is not enough to obtain reliable models. The data preprocessing stage is fundamental to improve the model's performance. Specifically, Feature Normalisation (FN) is commonly utilised to remove the features' magnitude aiming at equalising the features' contribution to the model training. Nevertheless, this work demonstrates that the FN method selection affects the model performance. Also, it is well-known that ANNs are commonly considered a ''black box'' due to their lack of interpretability. In this sense, several works aim to analyse the features' contribution to the network for estimating the output. However, these methods, specifically those based on network's weights, like Garson's or Yoon's methods, do not consider preprocessing factors, such as *dispersion factors*, previously employed to transform the input data. This work proposes a new features' relevance analysis method that includes the dispersion factors into the weight matrix analysis methods to infer each feature's actual contribution to the network output more precisely. Besides, in this work, the *Proportional Dispersion Weights (PWD)* are proposed as explanatory factors of similarity between models' performance results. The conclusions from this work improve the understanding of the features' contribution to the model that enhances the feature selection strategy, which is fundamental for reliably modelling a given problem.

INDEX TERMS Artificial neural networks, explainability, feature contribution, feature normalization.

I. INTRODUCTION

Artificial Neural Networks (ANNs) are algorithms that simulate the human brain learning behaviour, modelled by a weighted directed graph of interconnected nodes or neurons. These neurons are simple functions whose arguments are the weighted summation of the inputs to the node [1]. Due to their ability to solve challenging computational problems [2], [3], ANNs are widely applied in different fields, like industry among others [4]–[8]. However, they are still considered a ''black box'' since the network's predictions cannot be directly explained. Therefore, in the last decades, there has been a surge of interest in explainable Artificial Intelligence (xAI) approaches [9]. In this line, researchers have shown an increased claim in understating the features' contribution for modelling the network [10]–[12]. As authors in [13] expound, the goal of feature relevance explanation techniques is to describe the functioning of a model by measuring each feature's influence on the predicted output. Since feature relevance methods can be viewed as indirect techniques to explain a model, they have become a vibrant subject of study in the xAI field [14]–[18].

The understanding of the features' relevance is essential not only to explain the features' contribution to the model but also to conduct proper Feature Selection (FS) [19]–[21]. FS is traditionally considered a preprocessing technique. It is

The associate editor coordinating the review of this manuscript and approving it for publication was M. Venkateshkumar.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

well-known that in data analysis in general, and for ANN in particular, data preprocessing is one of the essential stages in the development of a solution, and the choice of preprocessing steps can often have a significant effect on the algorithm's performance [22]. In the era of digitalisation, hundreds of features from complex systems are usually monitored to extract valuable knowledge from the data [23]. In order to reduce the model complexity as well as to save memory and computational cost, features' relevance-based FS is commonly applied [24]–[26]. In some cases, the features' relevance calculation is conducted by means of network's weights-based feature importance analysis methods [27]–[29].

Along with FS, another commonly employed preprocessing approach is the linear normalisation of the input features. Feature Normalisation (FN) is often useful if the features present values that differ significantly in magnitude. Each FN method transforms a given dataset differently. The impact of the FN method's selection on the algorithm's performance has been experimentally studied by some researchers [30]–[33] to estimate the most appropriate one for a given problem. However, it remains the extended approach of employing the min-max normalisation method before the use of an ANN [34]–[38]. Despite the importance of data normalisation, no works are found that include the influence of data normalisation when analysing the features' contribution to the resultant ANN model.

Thus, this work advances the state-of-the-art by theoretically examining the impact of data normalisation on the relative contribution of the input features to the ANN and, ultimately, the algorithm's performance. For that purpose, this work presents a new proposal for feature's contribution analysis that extends Garson's and Yoon's methods to include the normalisation influence when estimating the features' contribution to the ANN. The theoretical conclusions are also experimentally validated.

Section II describes the ANN-based models and Section III presents the formulation of FN. In Section IV the Garson's and Yoon's traditional features' relevance analysis methods based on weight matrix analysis (Section IV-A) are presented, and Section IV-B describes a new proposal for the adaptation of these methods to include the dispersion factors in the computation of the feature's contribution. Section V describes the employed well-known datasets from UCI repository [39] and argues the proposed methods of this work. The experimental results of the analysis are collected in Section VI; and a discussion and proposal of future work are described in Section VII. Finally, Section VIII collects the conclusion of the work.

II. ARTIFICIAL NEURAL NETWORKS

An Artificial Neural Network is a weighted directed graph of interconnected neurons that propagates data from the input layer to the output layer by transforming such data to obtain valuable information for modelling a problem. A neuron receives the weighted values from the neurons of the previous layer. In the neuron, the sum of the weighted values is computed and employed as the argument of an activation function $\varphi : \mathbb{R} \longrightarrow \mathbb{R}$; being the identity $\varphi(x) = x$ the simplest one. The ANN architecture is flexible in the number of hidden layers and neurons per layer. The higher the number of hidden layers and the neurons that compose them, the higher the model complexity.

In this work, the network's layers are represented by $h \in \{0, 1, \ldots, H, H+1\}$, where H is the number of hidden layers, and $h = 0$ and $h = H+1$ symbolise the input and output layers, respectively. The number of neurons in the h-th hidden layer is denoted by n_h. Note that $n_0 = m$ is equal to the number of features, and for the single output problems, $n_{H+1} = 1$. The matrix weight of the edges that connect the neurons of the $h-1$ layer with the neurons of the h-th layer is $W^h \in \mathbb{R}^{n_{h-1} \times n_h}$, and b_h represents the bias of the h-th layer. Then, the mathematical formulation of an ANN-based model is:

$$Y = \varphi\left(\ldots \varphi\left(X \cdot W^{(1)} + b_1\right) \ldots W^{(H+1)} + b_{H+1}\right) \quad (1)$$

For $\varphi(x) = x$, (1) can be rewritten as

$$\hat{Y} = X \cdot \left(\prod_{h=1}^{H+1} W^h\right) + cte = X \cdot \mathbb{W} + cte. \quad (2)$$

For the single output problem, $\mathbb{W}$ is a vector of length m, where the entry $j \in \{1, \ldots, m\}$ represents the total weight the network assigns to the j-th feature.

From (1), and especially, when the activation function is the identity as in (2), the ANN's weights are the fundamental parameters that relate the input data with the estimated output. The ANN weights, along with the bias, are iteratively updated during the training phase of the model to obtain $\hat{Y} \approx Y$. However, for reaching so, not only the parameters training is determinant but also the quality of the input data. As authors in [40] remark, input data must be provided in the amount, structure and format that suits the data mining task. Besides, in order to avoid that the measurement unit affects the data mining task, all the features should be expressed in the same measurement units with a common scale or range. Feature Normalisation (FN) attempts to equalise the features' magnitude, and it is also employed to speed up the learning process in ANNs, helping the weights converge faster.

III. FEATURE NORMALIZATION

FN is a preprocessing technique widely employed to avoid the magnitude differences between the features of a given dataset. Any statistical-based FN method can be expressed as

$$\widetilde{X} = \frac{X - pos(X)}{dis(X)} \quad (3)$$

Equation (3) transforms a given dataset X into a normalised one $\widetilde{X}$ based on $pos(X)$ and $dis(X)$; $pos(X)$ refers to the position or central tendency statistic vector,[1] whereas

[1] For the sake of brevity, the vector composed by position or central tendency statistic is referred as position statistic from now on.

$dis(X)$ is the dispersion statistic vector which scales the features.

Equation (4) defines the decimal notation proposed to highlight the magnitude factors of each feature.

$$x_{ij} = sign(x_{ij})\, 0.d_1 d_2 d_3 \ldots \cdot 10^{n_j} = \widehat{x_{ij}} \cdot 10^{n_j} \tag{4}$$

In (4), $d_1, d_2, d_3, \ldots \in \{0, 1, \ldots, 9\}$ and $n_j \in \mathbb{Z}$ is fixed in such a way that $\forall j$, $|n_j|$ is the minimum number which fulfils: $X_j = \widehat{X_j} \cdot 10^{n_j}$, and $max|\widehat{X_j}| < 1$. Then, $\forall i, j$, $|\widehat{x_{ij}}| < 1$, and 10^{n_j} represents the magnitude factor of each feature. With the defined decimal notation, and since the statistical factors are estimated by linear operations, $pos(X_j) = pos(\widehat{X_j}) \cdot 10^{n_j}$ and $dis(X_j) = dis(\widehat{X_j}) \cdot 10^{n_j}$; FN can be re-written as

$$\widetilde{X_j} = \frac{\widehat{X_j} \cdot 10^{n_j} - pos(\widehat{X_j}) \cdot 10^{n_j}}{dis(\widehat{X_j}) \cdot 10^{n_j}} = \frac{\widehat{X_j} - pos(\widehat{X_j})}{dis(\widehat{X_j})} \tag{5}$$

Equation (5) shows that the magnitude factors in the normalised dataset disappear. This is the main reason why, as aforementioned in Section II, FN is widely employed to equalise the magnitude of the features. However, as a consequence of FN, each feature j is scaled by a dispersion factor $dis(\widehat{X_j})$ dependant on its values distribution.

Note that the normalised features present a dispersion equal to 1 in terms of the dispersion factor employed to transform the dataset. But, in order to fulfil $dis(\widetilde{X_j}) = 1$ each feature $\widehat{X_j}$ is differently expanded or compressed. Thus, the higher the value of $dis(\widehat{X_j})$, the higher the level of compression a feature undergoes, and consequently, the lower the contribution weight on the ML algorithm. Analogously, the lower the value of $dis(\widehat{X_j})$, the higher the expansion of $\widehat{X_j}$, and the higher the expected contribution to the model. Thus, the inverse of the dispersion factors can be viewed as unsupervised feature weights. In fact, in the ANN's first layer, the network's weights are multiplied by the normalisation weights, so the first layer's resulting weights are $dis(\widehat{X_j})^{-1} \cdot W_j^0$ $\forall j \in \{1, \ldots, m\}$. Then, since the dispersion factors act as weights along with the network's weights, it conditions the model performance and the features' contribution to the model.

IV. FEATURE RELEVANCE ANALYSIS METHODS

ANN-based models are considered a "black box" since the network's predictions cannot be directly explained. Therefore, several approaches to understand the features' contribution for modelling the network have been proposed. Some features' relevance analyses for ANN-based problems rely on the network's Weights Matrix Analysis (WMA) to estimate the features' contribution to the model. In this Section, first, the well-known Garson's and Yoon's methods are described. Next, a novel approach that considers the network's weights and the dispersion factors is proposed.

A. FEATURE RELEVANCE ANALYSIS METHODS BASED ON NETWORK's WEIGHT MATRIX

In order to understand the features' contribution to the model, WMA methods are usually employed. These methods, which belong to the features' relevance explanation techniques, calculate the features' contribution based on the network's weights related to each feature. Among the WMA methods, Garson's [41], and Yoon's methods [42] are well-known. They compute the features' contribution values as defined in (6) and (7), respectively.

$$Garson_j = \frac{|\prod_{h=1}^{H+1} W^h|_j}{\sum_{j=1}^{m} |\prod_{h=1}^{H+1} W^h|} = \frac{|\mathbb{W}_j|}{\sum_{j=1}^{m} |\mathbb{W}_j|} \in [0, 1] \tag{6}$$

$$Yoon_j = \frac{(\prod_{h=1}^{H+1} W^h)_j}{\sum_{j=1}^{m} |\prod_{h=1}^{H+1} W^h|} = \frac{\mathbb{W}_j}{\sum_{j=1}^{m} |\mathbb{W}_j|} \in [-1, 1] \tag{7}$$

Similarly to other features' relevance explanation techniques, it is considered that the higher the $Garson_j$ or $Yoon_j$ value is, the higher the features' contribution to the network.

Note that the preprocessed features implicitly influence the contribution values estimated by Garson's and Yoon's methods in the sense that the weights have been obtained from the training process with the preprocessed features (and not the raw features). In order to calculate more precisely the real feature's contribution to the model, a novel method that explicitly and formally considers the dispersion factors in addition to the network's weights is presented.

B. FEATURE RELEVANCE ANALYSIS METHODS BASED ON NETWORK's WEIGHT MATRIX AND DISPERSION FACTORS: A NEW PROPOSAL FOR THE ADAPTATION OF GARSON's AND YOON's METHODS

Despite data preprocessing –and hence FN– is considered essential to obtain quality results, until the date, no works that analyse the preprocessing stage impact for estimating the features' influence on the ANN-based model are found. This work aims to advance the state-of-the-art by including the dispersion factors in the features' contribution estimation.

Equation (2) can be viewed as the formula for $\hat{Y}$ estimation given a dataset X. However, before ANN employment, $\forall j \in \{1, \ldots, m\} X_j$ is usually transformed by a statistical-based normalisation method. Then, from (2) and (5), the mathematical formulation of an ANN-based model trained with a normalised dataset $\widetilde{X}$ can be re-defined as:

$$\hat{Y} = \widetilde{X} \cdot \left(\prod_{h=1}^{H} W^h \right) + cte = X \cdot \mathbb{D} \cdot \left(\prod_{h=1}^{H} W^h \right) + cte \tag{8}$$

where $\mathbb{D} = \text{diag}\ dis(X_1)^{-1}, \cdots, dis(X_m)^{-1}$ is the diagonal matrix, and the elements $\mathbb{D}_{jj}$ correspond to the inverse of the dispersion factor of the j-th feature. Equation (8) illustrates that the dispersion factors, in addition to the weight matrix, influence the features' contribution to the model. Consequently, in order to estimate the true impact of a given feature on the model, this work proposes to include the dispersion factors in Garson's and Yoon's methods for the features' influence calculations as follows:

$$\widehat{Garson_j} = \frac{(\mathbb{D} \cdot |\mathbb{W}|)_j}{\sum_{j=1}^{m} |(\mathbb{D} \cdot \mathbb{W})_j|} \in [0, 1] \tag{9}$$

$$\widehat{Yoon}_j = \frac{(\mathbb{D}\cdot\mathbb{W})_j}{\sum_{j=1}^{m} |(\mathbb{D}\cdot\mathbb{W})_j|} \in [-1, 1] \tag{10}$$

As described in Section IV-A, the higher the value of $\mathbb{W}_j$, the higher the j-th feature's contribution to the network. Similarly, by interpreting the inverse of the dispersion as unsupervised weights, as stated in Section III, the higher the value of $dis(X_j)^{-1}$, the higher the contribution of such feature to the model. Thus, the same rationale can be applied to $dis(X_j)^{-1} \cdot |\mathbb{W}_j|$.

V. MATERIALS AND METHODS

This Section describes the procedure employed to experimentally analyse and validate that the FN method selection influences the ANN-based model performance and hence, justify the inclusion of the dispersion factors in the features' contribution estimation. Given the experimental analysis and validation, Fig. 1 shows the high-level diagram of 1) the data split and preprocessing and 2) the ANN-based model training and evaluation conducted in this work. The elements employed in the following Sections to evaluate the FN influence on the ANN-based models are highlighted with a magnifying glass symbol. Note that Section III demonstrates that the magnitude factors disappear when normalising the features. Consequently, from now on $\widehat{X}$ (4) is employed.

A. DATASETS

In order to validate the hypothesis presented in Section IV-B, four public available real use cases from UCI repository [39] are employed.

TABLE 1. Description of datasets from UCI repository utilized in this work.

Dataset	Features	Samples	Y	Use case
CBM [43]	14	11934	[0.975, 1]	Naval propulsion plants-compressor decay state.
NOX [44] CO [44]	9	7384	[25.905, 119.68] [0.2128, 41.097]	Gas turbine CO and NOx emission. Data from 2015
News [45]	59	39644	[8, 731]	News online popularity

Table 1 summarises the utilised datasets. This work is focused on regression problems; then, the four datasets have continuous output values. Both NOX and CO datasets utilise the same input data. However, NOX dataset aims at estimating the Nitrogen oxides (NOx) emission from a gas turbine, while for CO the Carbon monoxide (CO) emission of the same gas turbine is registered.

B. DATA PREPARATION, ANN-BASED MODEL TRAINING, AND EVALUATION METRICS

The first step consists in preparing a given dataset to obtain the train/test subsets and normalise the data. Then, the ANN-based model is trained with the preprocessed data. Each step of Fig. 1 and the primary metrics employed to analyse the obtained results are next described.

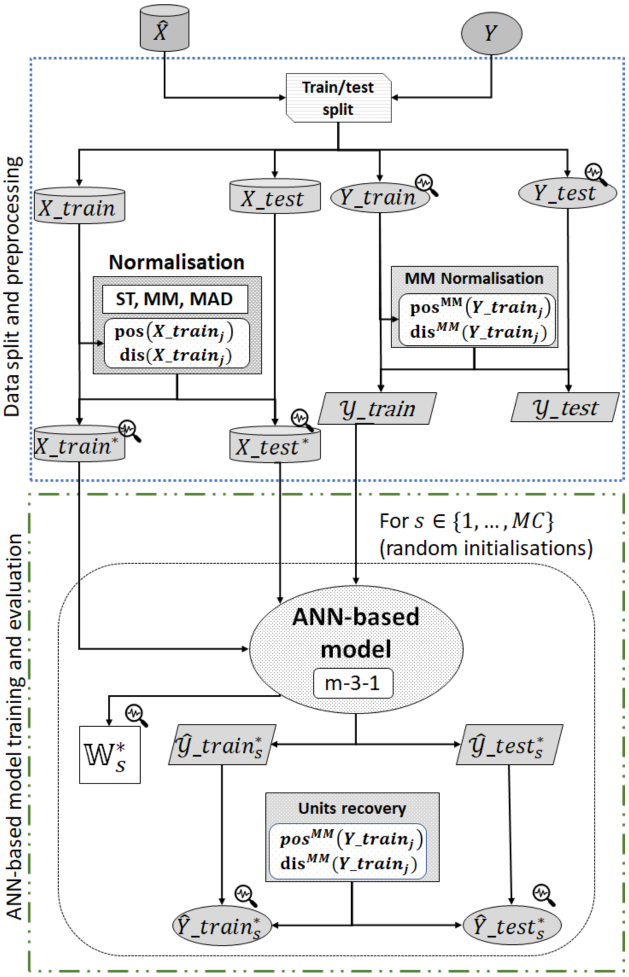


FIGURE 1. High-level diagram of the proposed method for data preprocessing, and ANN-based model train and evaluation.

1) DATASET SPLIT INTO TRAIN AND TEST SETS

A given dataset $X \in \mathbb{R}^{n\times m}$ composed by n samples described by m features and the associated real labels $Y \in \mathbb{R}^n$ are split into train (X_train, Y_train) and test (X_test, Y_test) disjoint sets of $n_{train} = 0.7 \cdot n$, and $n_{test} = n - n_{train}$ samples, respectively. The training set is employed to adjust the model's parameters (weights and bias), while the test set is utilised to validate the model's performance.

2) NORMALIZATION METHODS

In order to validate the impact of the FN method selection on the network, three well-known normalisation methods are employed in this work. Table 2 presents the selected normalisation methods and the statistical position and dispersion factors utilised to transform the features.

Each normalisation method from Table 2 utilises different position and dispersion statistics to transform the features of a given dataset. More concretely, ST, MM and MAD compress or expand each feature based on its standard deviation σ, range, and median absolute deviation *mad* dispersion statistics, respectively. Thus, ST and MAD calculate

TABLE 2. Normalization methods selected in this work for the analysis and validation of the proposal.

Normalisation method	$\mathbf{pos(X_j)}$	$\mathbf{dis(X_j)}$
Standardisation (ST)	$\overline{X_j}$	σ_j
Min-max (MM)	$\min(X_j)$	$range(X_j) = \max(X_j) - \min(X_j)$
Median Absolute Deviation (MAD)	$\mathrm{Me}(X_j)$	$mad(X_j) = \mathrm{Me}\left(\lvert X_j - \mathrm{Me}(X_j)\rvert\right)$

the dispersion of the features' samples around the mean and median values, respectively. In contrast, MM computes the statistical factors considering the extreme values of the features.

The FN methods from Table 2 are employed as follows: for each normalisation method $* \in \{ST, MM, MAD\} = Norm$ the statistical factors are calculated from X_train as described in (3). Then, they are applied to X_train and X_test to create train X_train^* and test X_test^* datasets. Thus, from a given dataset X, a normalised dataset $\widetilde{X}^*$ is obtained for each $* \in Norm$.

Independently of the FN method utilised to transform the input features, the output label is normalised with MM. The only difference between the analysed models is the normalisation method utilised for the input data transformation. Then, $min(Y\_train)$ and $range(Y\_train)$ values are utilised in (3) to calculate $\mathcal{Y}\_train$, $\mathcal{Y}\_test$.

3) ANN TRAINING STRATEGY FOR THE ANALYSIS OF THE NORMALISATION INFLUENCE

In this work, for each normalised dataset, an ANN with one hidden layer composed of three hidden neurons ($[m{-}3{-}1]$) is utilised. The neurons of the hidden and output layers are activated with the identity function. A maximum of 300 iterations is set, and the training stops if no improvement is observed for 10 iterations. The network's weights are initialised with Xavier's method [46], and $MC = 50$ random initialisations are utilised for each normalised dataset. In this way, for each initialisation, $s \in \{1, \ldots, MC\}$, the networks trained with each $\widetilde{X}^*$ employ the same initial network's weights.

The ANN is trained with X_train^* searching for the optimal weights and bias values that obtain, for each initialisation, $\hat{\mathcal{Y}}\_train^*_s \approx \mathcal{Y}\_train$. From each trained network, the estimated outputs $\hat{\mathcal{Y}}\_train^*_s$ and $\hat{\mathcal{Y}}\_test^*_s$ are calculated and re-scaled with the statistical factors of Y_train into the original units, obtaining $\hat{Y}\_train^*_s$ and $\hat{Y}\_test^*_s$. Besides, $\forall s$, the network's weight vector $\mathbb{W}_s$ is saved for further analyses.

4) METRICS FOR INFERRING THE NORMALISATION INFLUENCE

The main goal of this work is to validate the impact of the FN method selection influence on the model's performance and the adequacy of employing dispersion factors, in addition to the network's weight vector, to infer the features' contribution appropriately. For doing so, (1) the estimated outputs $\hat{Y}\_train^*_s$ and $\hat{Y}\_test^*_s$; (2) the statistical dispersion factors $dis(X_j)$; and (3) the weight vector $\mathbb{W}^*_s$ obtained from the trained models are analysed primarily based on the following metrics.

- Kendall's τ correlation coefficient [47] measures the degree of similarity between two ranks assigned to the same set of objects, i.e. paired rankings. Kendall's τ ranges from -1 to 1. A $\tau = 0$ indicates the non-relationship between the two rankings. If $\tau = -1$, a ranking is the inverse of the other, while $\tau = 1$ when both rankings are the same. Then, the higher the value of τ, the higher the ranks similarity.

- Distance is a key concept in many statistical and pattern recognition methods which measures the closeness or similarity between two objects. The Euclidean distance E_d between two vectors $\mathbf{a}, \mathbf{b}$ is defined as $E_d(\mathbf{a}, \mathbf{b}) = \sqrt{\sum_{j=1}^{m}(a_j - b_j)^2}$, and it is equal to 0 if the components of both vectors are the same. The higher E_d, the higher the dissimilarity between the components of the vectors. Although E_d is scale dependant, in this work, it is applied to vectors with components ranging from 0 to 1.

- Performance measures are utilised to analyse the error or the similarity between two output features $\mathbf{y_a}, \mathbf{y_b}$ of length n. In this work, the mean absolute error (MAE), the root mean squared error (RMSE) and the coefficient of determination (R^2) regression performance measures are employed. MAE$(\mathbf{y_a}, \mathbf{y_b}) = (1/n)\sum_{i=1}^{n} \lvert y_{a_i}, y_{b_i} \rvert$ and RMSE$(\mathbf{y_a}, \mathbf{y_b}) = (1/n)\sqrt{\sum_{i=1}^{n}(y_{a_i}, y_{b_i})^2}$ measures the error as the mean absolute and the mean square quadratic differences between the elements of both output features, respectively. Thus, the lower the MAE and RMSE values, the higher the similarity between $\mathbf{y_a}$ and $\mathbf{y_b}$. In contrast, R$^2(\mathbf{y_a}, \mathbf{y_b}) = 1 - \left(\sum_{i=1}^{m}(y_{a_i} - y_{b_i})^2 / \sum_{i=1}^{m}(y_{a_i} - \overline{y_a})^2\right)$ is a statistical measure of how well the regression predictions y_b approximate points of y_a. R^2 takes values up to 1. Values of R^2 lower than 0 appear when the model fits the data worse than a horizontal hyper-plane, while R$^2 = 1$ indicates that the regression predictions perfectly fit the data. Then, the higher the R^2, the higher the similarity between $\mathbf{y_a}$ and $\mathbf{y_b}$.

C. ANALYSIS OF THE NORMALIZATION INFLUENCE ON THE MODEL's PERFORMANCE

The first analysis aims to verify that the model's performance varies depending on the selected FN method. In particular, the analysis lies in 1) studying the differences between the outputs predicted by the models trained with the differently normalised datasets and 2) comparing the ANN-based models' performance depending on the FN method.

1) DIFFERENCE BETWEEN THE PREDICTIONS ESTIMATED BY THE DIFFERENT MODELS

In order to analyse the difference between the predictions estimated by the different models, for each $s \in \{1, \ldots, MC\}$ the MAE, RMSE and R^2 between $\hat{Y}\_train^*_s$ and $\hat{Y}\_train^+_s$ and between $\hat{Y}\_test^*_s$ and $\hat{Y}\_test^+_s$ for $* \neq + \in Norm$ are calculated, and the maximum, mean, minimum and standard deviation (std) values are computed for MC initialisations. If the selection of the FN method does not influence the model's performance, then MAE $=$ RMSE $= 0$ and R$^2 = 1$. Otherwise, differences between the estimated outputs would demonstrate the influence on the model's performance of the FN methods.

2) ANN PREDICTION PERFORMANCE DEPENDING ON THE FN METHOD

Complementary, in order to analyse the ANN prediction performance depending on the FN method, for each random initialisation, MAE, RMSE and R^2 between $\hat{Y}\_train_s^*$ and $\hat{Y}\_train$, and $\hat{Y}\_test_s^*$ and $\hat{Y}\_test$ are calculated together with the maximum, mean, minimum and std values estimated for the MC initialisations. Similarly, if the normalisation method selection does not affect the model's performance, the same MAE, RMSE and R^2 statistical values are expected independently from the FN method employed when transforming the data. In contrast, differences in the estimated performance would also demonstrate the hypothesis of this work.

Complementary, the non-parametric Wilcoxon signed-rank test [48] is employed to check the existence of statistical differences between $RMSE(Y\_train, Y\_train^*)$ and $RMSE(Y\_train, Y\_train^+)$, or between $RMSE(Y\_test, Y\_test^*)$ and $RMSE(Y\_test, Y\_test^+)$ for $* \neq + \in Norm$. In Wilcoxon signed-rank test, the same subjects are evaluated under two different conditions. In this case, each subject is the model with the s-th random initialisation of the weights, and the different conditions are the FN methods $* \neq + \in Norm$ utilised to transform the network's input features. The null hypothesis H_0 of Wilcoxon signed-rank test assumes that the related samples $[RMSE(Y\_train, Y\_train_1^*), \ldots, RMSE(Y\_train, Y\_train_{MC}^*)]$ and $[RMSE(Y\_train, Y\_train_1^+), \ldots, RMSE(Y\_train, Y\_train_{MC}^+)]$ come from the same population, i.e, the distribution of differences has a median of zero. The test's p-values are calculated and, if p-value< 0.05, H_0 is rejected with a significant level of 5%.

D. ANALYSIS OF THE DISPERSION FACTORS AS EXPLANATORY FACTORS OF THE VARIATIONS IN MODEL's PERFORMANCE

Each FN method collected in Table 2 employs different dispersion statistics or factors to transform the input features. Then, as stated in Section III, it is expected that each FN method $* \in Norm$ transforms differently a given dataset, which ultimately conditions the features' contribution values and, consequently, the ANN-based model's performance. Once verified that FN methods impact the model's performance, the dispersion factors are analysed as explanatory factors of such variations. It is assumed that a relationship exists between the results in the ANN-based model performance and the dispersion factors. Thus, the higher the differences between the dispersion factors, the higher the difference between the output estimations and the weight vector of the models trained with different $\widetilde{X}^*$. The analysis of the dispersion factors as explanatory factors is conducted as follows:

1) ANALYSIS OF PROPORTIONAL DISPERSION FACTORS

In this work, first, $\forall * \in Norm$ the scaling *dispersion factors* $w_j^* = 1/dis^*(X_j)$, specifically, their *Proportional Dispersion Weight (PDW)* estimated as $\hat{w}_j^* = w_j^* / \sum_{j=1}^m w_j^*$ are analysed. In order to infer the expected similarity between $\widetilde{X}^*$ and $\widetilde{X}^+$ for $* \neq + \in Norm$, the Kendall's τ correlation and the Euclidean distance between $\hat{w}_j^*$ and $\hat{w}_j^+$ are calculated to evaluate the similarity between the PDWs employed to create the different normalised datasets.

2) SIMILARITY BETWEEN PDW AND PERFORMANCE RESULTS

Once estimated the PDWs for each normalisation method, the level of similarity between the dispersion factors are compared accordingly with the level of similarity between the model's performance reached from Section V-C2 by differently normalised datasets. The coherence between both will allow setting the dispersion factors as explanatory factors of the variations in the model's performance.

E. ANALYSIS OF THE NORMALISATION INFLUENCE ON THE FEATURES' CONTRIBUTION

As described in Section II, Garson's and Yoon's methods are based on the network's weights to estimate the contribution of each feature in the model. From (6) and (7) it is observed that the main difference in the resulting features' contribution values is due to the direction, so, for the sake of brevity, from now, only Garson's method is considered.

Thus, G^* and $\hat{G}^*$ represent the features' contribution values calculated with the traditional and the adapted Garson's methods, respectively. $* \in Norm$ refers to the FN method employed to obtain the X_train^*, so as the weight vectors $\mathbb{W}^*$ from (6) and $\mathbb{D}^* \cdot \mathbb{W}^*$ from (9) can be computed. Then, for each $*$, MC networks with different initial weights are trained, and G_s^* and $\hat{G}_s^*$ are finally computed.

Once analysed the FN method selection influence on the model's performance and the relationship between the PDWs and the estimated outputs, this Section studies the impact of FN on the features' contribution values and the adequacy of the proposed adapted Garson's method to calculate the real features' influence. For doing so, first, an analysis of the features' contribution values in terms of the traditional and the adapted Garson's method is conducted. Then, a comparison with the results from Sections V-C and V-D is performed. Finally, a Feature Selection strategy is applied in order to demonstrate the superiority of the proposed adapted Garson's method for estimating the real features' contribution.

1) MEAN FEATURES CONTRIBUTION

In order to analyse the FN method selection impact on the features' influence on the network, the mean features' contribution values resulting from the MC random initialisation based on the traditional $\overline{G}^* = (1/MC)\sum_{s=1}^{MC} G_s^*$ and on the proposed adapted Garson's method $\overline{\hat{G}}^*$ are calculated and analysed considering the steps described below.

a: TRADITIONAL GARSON's METHOD

First, the differences between the weight matrix-based features contribution derived from the selection of the FN method is analysed. In order to inspect the $\overline{G}_j^*$ values distribution and the discriminative influence of the j-th feature: 1) the difference between the maximum and the minimum, and 2) the standard deviation of the features' contribution values are calculated. Then, aiming at examining the effect of FN in the features' influence on the model, a pairwise comparison between $\overline{G}^*$ and $\overline{G}^+$ with $* \neq +$ is conducted in terms of Kendall's τ correlation coefficient and Euclidean distance.

b: PROPOSED ADAPTED GARSON's METHOD

The same analysis is performed over $\hat{\overline{G}}^*$ to inspect the features' contribution computed with the adapted Garson's method.

c: COMPARISON BETWEEN THE TRADITIONAL AND THE PROPOSED ADAPTED GARSON's METHOD

Finally, with the aim of inferring the validity of the proposed adapted Garson's method to estimate the real features' contribution, first, a comparison between $\overline{G}^*$ and $\hat{\overline{G}}^*$ is performed. Then, the results from Sections V-C2 and V-D are here utilised to infer from the correspondence between the models' performance, the dispersion factors and the features' relevance analysis methods the superiority of the proposed adapted Garson's method.

2) FEATURE SELECTION BASED ON THE FEATURES' CONTRIBUTION

In order to demonstrate the superiority of the proposed adapted Garson's method, a FS strategy is conducted to analyse the effect of removing features considering the traditional Garson's method versus the proposed adapted one. The estimated features' contribution values from the models that obtain the lowest RMSE are employed for this strategy. Then, for each $* \in Norm$, the feature' influence values calculated with the traditional Garson's method are denoted as $\underline{G}^*$, while the estimated with the proposed one are referred to as $\hat{\underline{G}}^*$. The FS based on the features' contribution values computed with the traditional or the proposed Garson's methods ($f_C \in \{\underline{G}^*, \hat{\underline{G}}^*\}$) is applied as described in Algorithm 1.

VI. EXPERIMENTAL VALIDATION

This Section shows the experimental results obtained from training and testing the ANN architecture presented in Section V-B3. More concretely, first, the influence of the FN methods on the models' performance is studied. Next, an analysis of the proportional dispersion weights as explanatory factors of the estimated outputs is presented. Finally, the impact of FN on the features' contribution is demonstrated, and the superiority of the proposed adapted Garson's method is validated.

Algorithm 1 Feature Selection Strategy

```
for $f_C \in \{\underline{G}^*, \hat{\underline{G}}^*\}$ do
    for $ite \in \{1, \dots, m-1\}$ do
        Remove the $ite$ features with lowest $f_C$ value.
        Train the network, estimate the output and re-scale it to the original units.
        Estimate the RMSE between the real labels and the estimated ones.
    end for
end for
Plot the RMSE values estimated based on $\underline{G}^*$, and $\hat{\underline{G}}^*$ jointly with the RMSE estimated with all the features to analyse the effect of the feature removal.
```

A. ANALYSIS OF THE NORMALISATION INFLUENCE ON THE MODEL's PERFORMANCE

As described in Section V-C, an analysis of the dissimilarity between the outputs estimated by the models trained with differently normalised datasets is conducted. Note that $\forall * \in Norm$, the same 50 random initialisations establish the initial weights of the ANN. Thus, the only differences when training the models are the FN methods utilised to transform the input features.

1) DIFFERENCE BETWEEN THE PREDICTIONS ESTIMATED BY THE DIFFERENT MODELS

First, the comparison between the estimated outputs obtained from the differently normalised datasets is conducted.

Table 3 collects for each dataset the maximum, mean, minimum and standard deviation of MAE, RMSE and R^2 values from comparing the estimated $\hat{Y}\_train^*$ with $\hat{Y}\_train^+$ and $\hat{Y}\_test^*$ with $\hat{Y}\_test^+$ for $* \neq + \in Norm$. Given that similar results are obtained from train and test sets, for sake of brevity only the results from the training set are described. From the calculated scores presented in Tables 3a to 3d variations in the estimated outputs derived from the FN method selection can be inferred. In News, NOX and CO datasets the mean MAE is up to 1021.831, 0.363 and 1.547, respectively. Similarly, the mean $RMSE(\hat{Y}\_train^*, \hat{Y}\_train^+)$ values obtained are higher than 0.28, 1.21, and 86.41, respectively. In the case of CBM dataset, the RMSE vales are close to zero. However, in Table 3a, the mean R^2 values are lower than 0.605 when comparing Y_train^{MM} with Y_train^{ST} or Y_train^{MAD}, respectively. Then, from Table 3 it is concluded that the predictions considerably vary depending on the FN method selected for the feature preprocessing phase.

2) COMPARISON BETWEEN THE MODELS' PERFORMANCE SCORES

As demonstrated above, different outputs are obtained from the models trained with differently normalised data. As an example, Fig. 2 depicts the Y_test and Y_test^* obtained for $* \in Norm$ from the NOX dataset.

TABLE 3. Comparison between the estimated outputs $\hat{Y}^{*}$, $\hat{Y}^{+}$ for $* \neq + \in$ *Norm* in terms of MAE, RMSE and R^2.

		Train			Test		
		ST vs MM	ST vs MAD	MM vs MAD	ST vs MM	ST vs MAD	MM vs MAD
MAE	max	0.005	0.003	0.005	0.005	0.003	0.005
	mean	**0.003**	0.000	**0.003**	**0.003**	0.000	**0.003**
	min	0.001	0.000	0.001	0.001	0.000	0.001
	std	0.001	0.001	0.001	0.001	0.001	0.001
RMSE	max	0.004	0.003	0.004	0.004	0.003	0.004
	mean	**0.002**	0.000	**0.002**	**0.002**	0.000	**0.002**
	min	0.001	0.000	0.001	0.001	0.000	0.001
	std	0.001	0.000	0.001	0.001	0.000	0.001
R^2	max	0.874	0.999	0.822	0.873	0.999	0.821
	mean	**0.605**	0.967	**−0.324**	**0.604**	0.967	**−0.345**
	min	−0.040	−0.018	−7.594	−0.066	−0.053	−8.070
	std	0.163	0.142	1.320	0.165	0.147	1.382

(a) CBM turbine

		Train			Test		
		ST vs MM	ST vs MAD	MM vs MAD	ST vs MM	ST vs MAD	MM vs MAD
MAE	max	1032.526	3199.007	3226.765	900.803	845.512	986.802
	mean	70.406	**1021.831**	1017.157	60.408	**106.492**	91.797
	min	12.986	17.183	18.070	12.979	13.049	12.178
	std	185.677	832.802	830.036	165.592	150.734	136.101
RMSE	max	831.934	643.061	715.432	828.236	629.392	709.561
	mean	48.247	**86.412**	74.341	48.018	**80.288**	68.274
	min	9.553	9.990	9.112	9.474	10.029	9.087
	std	141.181	114.295	98.760	140.604	112.828	97.985
R^2	max	0.993	0.988	0.986	0.993	0.993	0.994
	mean	0.872	**−64.589**	**−71.387**	0.831	**0.597**	**−0.176**
	min	−1.687	−430.938	−441.726	−4.272	−3.645	−41.970
	std	0.433	99.112	102.614	0.764	0.761	5.992

(b) News

		Train			Test		
		ST vs MM	ST vs MAD	MM vs MAD	ST vs MM	ST vs MAD	MM vs MAD
MAE	max	3.308	1.532	3.319	3.291	1.541	3.359
	mean	1.428	0.428	**1.547**	1.439	0.429	**1.559**
	min	0.482	0.140	0.475	0.479	0.139	0.478
	std	0.701	0.224	0.741	0.706	0.227	0.746
RMSE	max	2.614	1.232	2.633	2.614	1.229	2.668
	mean	1.117	0.339	**1.213**	1.123	0.339	**1.220**
	min	0.382	0.116	0.383	0.379	0.115	0.380
	std	0.547	0.179	0.581	0.551	0.180	0.585
R^2	max	0.997	1.000	0.997	0.997	1.000	0.997
	mean	0.966	0.997	0.957	0.964	0.997	0.956
	min	0.852	0.969	0.826	0.850	0.968	0.824
	std	0.035	0.005	0.042	0.036	0.005	0.043

(c) NOX

		Train			Test		
		ST vs MM	ST vs MAD	MM vs MAD	ST vs MM	ST vs MAD	MM vs MAD
MAE	max	0.859	0.372	1.068	0.873	0.363	1.079
	mean	0.327	0.110	**0.363**	0.327	0.110	**0.364**
	min	0.084	0.029	0.080	0.082	0.029	0.077
	std	0.177	0.060	0.191	0.176	0.060	0.190
RMSE	max	0.657	0.290	0.833	0.666	0.279	0.833
	mean	0.255	0.086	**0.284**	0.254	0.086	**0.284**
	min	0.065	0.023	0.061	0.065	0.023	0.060
	std	0.138	0.047	0.150	0.136	0.046	0.148
R^2	max	0.998	1.000	0.998	0.998	1.000	0.998
	mean	0.959	0.995	0.946	0.958	0.995	0.944
	min	0.802	0.966	0.622	0.803	0.966	0.600
	std	0.044	0.005	0.066	0.044	0.005	0.067

(d) CO

As Fig. 2 shows, the Y_test^{*} values do not match the real labels, and their values considerably differ depending on the FN method. For instance, in the zoomed subplot for sample number 1813, the estimated output for MM is more than 3 units lower than the estimated with ST and MAD; so differences in the performance of the models trained with the different normalised sets are expected.

Next, the model's performance of each selected dataset is analysed as aforementioned in Section V-C. Table 4 collects for $* \in$ *Norm* the maximum, mean, minimum and standard deviation of MAE, RMSE and R^2 values calculated for $\hat{Y}\_train_s^{*}$ with respect to Y_train, and for $\hat{Y}\_test^{*}$ with respect to Y_test. Note that the models obtain similar performance results for train and test sets, and since this work does not aim to analyse the models' generalisation ability, only the results over the train set are described.

As inferred from Table 3, and as Table 4 shows, FN method selection affects the model's performance. For instance, for News dataset (Table 4b), depending on $* \in$ *Norm*, there is a difference up to 29.928 and 40.225 in terms of mean MAE and RMSE, respectively. For the rest of datasets regarding $*$, the differences in terms of mean MAE or RMSE are lower than 0.1. However, in the case of the CBM dataset, the 0.002 of increment in the error depending the FN method corresponds to 8% of the original range of the real output (Table 1). Nevertheless, although the models' performance differences in Tables 4a, 4c and 4d may not seem significant, notice that Table 3 shows considerable differences between the models' outputs. Then, in order to complement the conclusions derived from Table 4, Table 5 collects the p-values obtained with the Wilcoxon signed-rank test for assessing significant differences in the model's performance regarding the FN method with which the training and test sets are normalised.

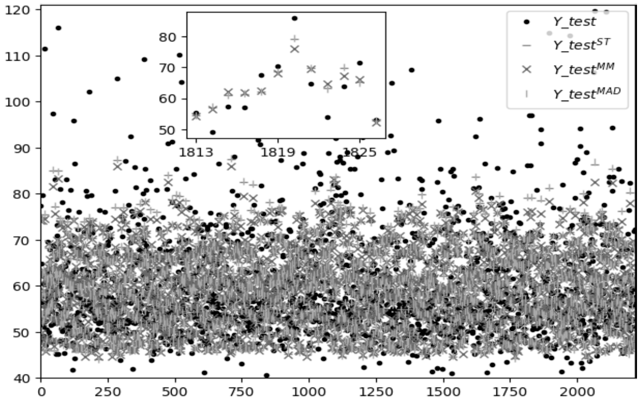


FIGURE 2. NOX.

The null hypothesis H_0, which states no statistical differences in the model's performance –in terms of RMSE– derived from the FN method selection, can be rejected in 17 out of 24 performed tests with a significance level of 5%. These 17 p-values, that represent the 70.833% of the p-values collected in Table 5, are remarked with bold text. In the rest of the cases (ST with respect to MM for News in the test set, and in both train and test sets for CBM in ST with respect to MAD, and MM with respect to ST and MAD of CO datasets), there is no evidence for rejecting H_0. However, Table 3b for the News dataset shows that the mean±std values of RMSE($\hat{Y}\_test^{ST}$, $\hat{Y}\_test^{MM}$) estimated from the 50 random initialisation is 48.018 ± 140.604 (more than 6% of the range of the real labels of the dataset in Table 1). Similarly, for train and test sets, the mean±std values depicted in Table 3d when comparing the RMSE of the outputs estimated for CO dataset normalised with MM with respect to ST or MAD are 0.25 ± 0.14 and 0.28 ± 0.15, respectively (around 1% of the range of the real labels in Table 1). Then, although in the mentioned cases there is no evidence for rejecting H_0, with the calculated statistics, significant differences are inferred when the estimated outputs obtained by different $*$ are straightly compared.

All in all, it can be concluded that the selection of a normalisation method for the preprocessing phase results in significant differences in the model's performance.

TABLE 4. Maximum, mean, minimum and standard deviation of MAE, RMSE and R^2 values for comparing the real label Y and the estimated one $\hat{Y}^*$ for each of the 50 random initializations and for each $* \in$ *Norm*.

		Train				Test			
		raw	ST	MM	MAD	raw	ST	MM	MAD
MAE	max	0.007	0.005	0.007	0.005	0.007	0.005	0.007	0.005
	mean	0.006	0.004	0.006	0.004	0.006	0.004	0.006	0.004
	min	0.006	0.002	0.004	0.002	0.006	0.002	0.004	0.002
	std	0.000	0.001	0.000	0.001	0.000	0.001	0.000	0.001
RMSE	max	0.008	0.006	0.008	0.006	0.008	0.006	0.008	0.006
	mean	0.007	0.005	0.007	0.005	0.007	0.005	0.007	0.005
	min	0.007	0.003	0.005	0.003	0.007	0.003	0.005	0.003
	std	0.000	0.001	0.000	0.001	0.000	0.001	0.000	0.001
R^2	max	0.079	0.843	0.494	0.850	0.082	0.849	0.498	0.857
	mean	0.025	0.583	0.235	0.589	0.026	0.588	0.235	0.594
	min	-0.016	0.352	-0.010	0.347	-0.016	0.354	-0.020	0.348
	std	0.023	0.120	0.080	0.132	0.025	0.122	0.082	0.134

(a) CBM turbine

		Train				Test			
		raw	ST	MM	MAD	raw	ST	MM	MAD
MAE	max	127.948	831.368	126.486	726.563	127.543	829.684	126.386	721.834
	mean	126.346	153.708	126.027	155.955	126.241	153.447	126.042	150.022
	min	125.809	124.746	125.769	125.405	125.749	124.436	125.731	124.715
	std	0.362	117.345	0.166	84.068	0.297	117.045	0.181	83.712
RMSE	max	150.145	1039.799	149.107	3230.031	149.874	914.020	149.214	1000.068
	mean	148.641	188.782	148.527	1045.089	148.768	182.990	148.779	184.921
	min	148.311	148.387	148.272	148.611	148.400	147.293	148.383	147.824
	std	0.361	159.569	0.172	808.528	0.288	138.917	0.177	119.457
R^2	max	0.522	0.521	0.522	0.520	0.517	0.524	0.517	0.521
	mean	0.520	−0.329	0.520	−36.969	0.515	−0.158	0.514	−0.063
	min	0.510	−22.513	0.516	−225.889	0.507	−17.326	0.512	−20.939
	std	0.002	3.686	0.001	53.142	0.002	2.895	0.001	2.992

(b) News

		Train				Test			
		raw	ST	MM	MAD	raw	ST	MM	MAD
MAE	max	5.900	6.072	6.038	5.985	5.857	6.037	6.039	5.971
	mean	5.562	5.493	5.535	5.512	5.542	5.419	5.474	5.441
	min	5.476	5.348	5.407	5.351	5.453	5.281	5.332	5.252
	std	0.074	0.115	0.124	0.124	0.073	0.120	0.141	0.136
RMSE	max	7.836	8.043	8.105	7.875	7.728	8.044	8.099	7.819
	mean	7.637	7.333	7.428	7.356	7.555	7.260	7.363	7.282
	min	7.520	7.131	7.241	7.136	7.455	7.062	7.171	7.063
	std	0.065	0.152	0.191	0.156	0.057	0.159	0.197	0.163
R^2	max	0.554	0.599	0.586	0.598	0.526	0.575	0.562	0.575
	mean	0.540	0.575	0.564	0.573	0.513	0.550	0.538	0.548
	min	0.515	0.489	0.482	0.511	0.491	0.448	0.441	0.479
	std	0.008	0.018	0.023	0.018	0.007	0.020	0.026	0.021

(c) NOX

		Train				Test			
		raw	ST	MM	MAD	raw	ST	MM	MAD
MAE	max	1.026	1.063	0.915	1.178	1.009	1.047	0.903	1.156
	mean	0.956	0.783	0.772	0.797	0.939	0.777	0.766	0.791
	min	0.876	0.736	0.738	0.737	0.864	0.730	0.731	0.731
	std	0.034	0.052	0.040	0.069	0.032	0.051	0.040	0.068
RMSE	max	1.676	1.593	1.480	1.703	1.876	1.820	1.710	1.913
	mean	1.596	1.294	1.291	1.306	1.808	1.546	1.543	1.557
	min	1.500	1.256	1.257	1.256	1.729	1.512	1.514	1.512
	std	0.042	0.052	0.049	0.069	0.036	0.047	0.043	0.061
R^2	max	0.533	0.672	0.672	0.673	0.447	0.577	0.576	0.577
	mean	0.471	0.652	0.654	0.645	0.395	0.558	0.559	0.551
	min	0.417	0.473	0.545	0.398	0.349	0.387	0.459	0.323
	std	0.028	0.030	0.027	0.041	0.024	0.028	0.025	0.038

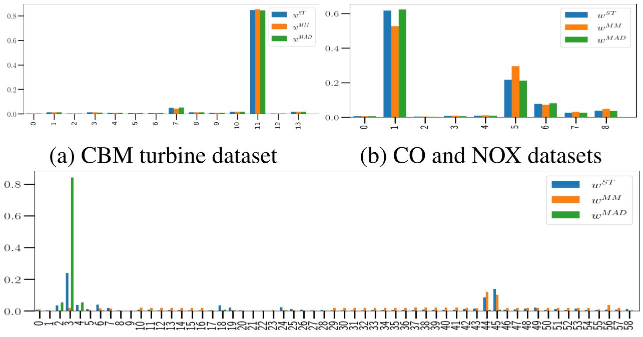
(d) CO

TABLE 5. P-values obtained from Wilcoxon signed-rank test.

	Train			Test		
	ST vs MM	ST vs MAD	MM vs MAD	ST vs MM	ST vs MAD	MM vs MAD
CBM	**0.000**	0.124	**0.000**	**0.000**	0.124	**0.000**
News	**0.000**	**0.000**	**0.000**	0.075	**0.000**	**0.000**
NOX	**0.001**	**0.005**	**0.021**	**0.001**	**0.010**	**0.017**
CO	0.559	**0.000**	0.110	0.633	**0.000**	0.121

B. ANALYSIS OF THE DISPERSION FACTORS AS EXPLANATORY FACTORS

As explained in Section II the network's weights adjust the features' contribution in order to create a model that estimates $\hat{Y} \approx Y$. Nevertheless, the hypothesis of this work is that the dispersion factors influence the model's training, and consequently, the model's performance. The former hypothesis has been validated in Section VI-A. In order to study the influence of the FN method selection, first, an analysis and comparison of the proportional dispersion weights estimated by different FN methods are conducted. Then, an analysis of the similarity of these factors and the output estimations over the differently normalised datasets is performed.



(a) CBM turbine dataset

(b) CO and NOX datasets

(c) News datasets

FIGURE 3. Proportional dispersion weights (PDW) $\hat{w}^*$ for $* \in$ *Norm*.

1) ANALYSIS OF THE PROPORTIONAL DISPERSION FACTORS

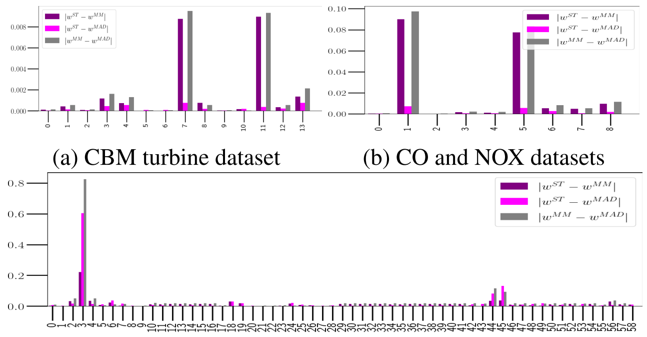
Fig. 3 shows for each dataset and for $* \in Norm$, the proportional dispersion weights $\hat{w}_j^*$ estimated for each feature. In Fig. 3c for News dataset, especially in features 2, 3, 4, 44 and 45, it is observed that $\hat{w}_j^*$ significantly differs depending on the normalisation method employed. In fact, in News dataset, $\forall * \in Norm$, feature 3 obtains the highest PDW, but $\hat{w}_3^{ST}$ takes values closer to $\hat{w}_3^{MM}$ than to $\hat{w}_3^{MAD}$. In contrast, for CBM, NOX and CO datasets, $\hat{w}_j^{ST}$ and $\hat{w}_j^{MAD}$ present the most similar values.

In order to conduct the pairwise comparison between the dispersion factors, Fig. 4 depicts the absolute difference between $\hat{w}_j^*$ and $\hat{w}_j^+$ for $* \neq + \in Norm$.

Fig. 4c clearly shows that in News dataset $|\hat{w}_j^{ST} - \hat{w}_j^{MM}| < |\hat{w}_j^{ST} - \hat{w}_j^{MAD}|$ for $j \in \{3, 44, 45\}$; while for $j \in \{6, 18, 19, 24, 58\}$ the minimum $|\hat{w}_j^* - \hat{w}_j^+|$ is reached with MM and MAD. In contrast, $\hat{w}^{ST}$ and $\hat{w}^{MAD}$ present the lowest absolute differences in Figs. 4a and 4b.

Table 6 describes the similarity between $\hat{w}_j^*$ and $\hat{w}_j^+$ for $* \neq + \in Norm$ in terms of Kendall's τ correlation and Euclidean distance.

For News dataset, $\tau(\hat{w}^{ST}, \hat{w}^{MM})$ and $\tau(\hat{w}^{MM}, \hat{w}^{MAD})$ values from Table 6a are far from 1, demonstrating that each feature's position in the rank derived from $\hat{w}^*$ significantly varies depending on $* \in Norm$. When comparing $\hat{w}^{MM}$ with $\hat{w}^{ST}$ or $\hat{w}^{MAD}$ in CBM dataset, or $\hat{w}^{ST}$ with $\hat{w}^{MAD}$ in News dataset, τ ranges between 0.818 and 0.889. So, minor PDWs' rank variations can be found depending on $* \in Norm$. The only case in which the proportional dispersion weights' ranking does not vary depending on $*$ is observed in NOX or CO datasets. However, if the Euclidean distance between PDW values is analysed, it can be concluded that there are differences between the proportional estimated values with ST or MAD respect to the obtained with MM, from which differences in the network's performance can be foreseen.



(a) CBM turbine dataset

(b) CO and NOX datasets

(c) News datasets

FIGURE 4. **Absolute differences between the proportional dispersion weights $|\hat{w}^* - \hat{w}^+|$ for $* \neq + \in$ *Norm*.**

TABLE 6. **Similitude analysis between the proportional dispersion weights $\hat{w}^*, \hat{w}^+$ for $* \neq + \in$ *Norm*.**

dataset	ST vs MM	ST vs MAD	MM vs MAD
CBM	0.889	0.978	0.867
News	0.116	0.818	−0.046
NOX-CO	1	1	1

(a) Kendall's $\tau(\hat{w}^*, \hat{w}^+)$.

dataset	ST vs MM	ST vs MAD	MM vs MAD
CBM	0.02	0.01	0.029
News	0.147	0.042	0.173
NOX-CO	0.023	0.009	0.031

(b) $E_d(\hat{w}^*, \hat{w}^+)$.

2) SIMILARITY BETWEEN PDW AND PERFORMANCE RESULTS

Next, an analysis of $\hat{w}^*$ as explanatory factors of the similarities between the model's performance obtained by the dataset normalised by different FN methods is conducted.

According to Fig. 3c and Table 6, for News dataset, $\hat{w}^{ST}$ and $\hat{w}^{MM}$ are the most similar PDWs. These results match with those from Table 3b were the lowest MAE and RMSE and highest R^2 result from juxtaposing $\hat{Y}\_train^{ST}$ with $\hat{Y}\_train^{MM}$; while the mean R^2 value obtained when comparing MAD with ST or MM is lower than −64.589. Besides, Tables 6a and 6b show that the lowest τ and the highest E_d values are obtained when examining the PDW of News datasets. Similarly, Table 4b presents the highest differences between the model's performance resulting from the different FN methods (up to 40.225 and 34.211 in terms of mean RMSE).

Similarly, for CBM dataset, the Kendall's τ in Table 6 is lower than 0.9 when comparing the ranks of $\hat{w}^{ST}$ or $\hat{w}^{MAD}$ respect to $\hat{w}^{MM}$. Consequently, in Table 3a the mean R^2 is 0.6 and −0.324 for the mentioned cases, respectively.

In contrast, $\hat{w}^{ST}$ and $\hat{w}^{MAD}$ are the most similar PDWs for CBM, NOX and CO datasets (Fig. 3 and Table 6). Similarly, in Tables 3a, 3c and 3d the lowest mean MAE and RMSE values are obtained when comparing $\hat{Y}\_train^{ST}$ with $\hat{Y}\_train^{MAD}$. In fact, in NOX and CO datasets, the mean MAE and RMSE errors between the outputs estimated with MM respect to the calculated ones with ST or MAD are more than 3.2 times higher than the resulting from comparing ST and MAD. Besides, for these two datasets, $\tau(\hat{w}^*, \hat{w}^+) = 1$ (see Table 6), which explains that in Tables 3c and 3d the mean R$^2(\hat{Y}\_train^*, \hat{Y}\_train^+)$ are higher than 0.94 for $* \neq + \in$ *Norm*.

All in all, it is demonstrated that the higher the similarity between $\hat{w}^*$ and $\hat{w}^+$ for $* \neq + \in$ *Norm*, the lower the difference expected between the output estimations resulting from the dataset normalised with $*$ and $+$. Thus, in order to select among different FN methods the suitable one for the problem at hand, by knowing in advance the similarity between $\hat{w}^*$ and $\hat{w}^+$, the expected similarity between Y_train^* and Y_train^+ can be inferred.

C. ANALYSIS OF THE NORMALIZATION INFLUENCE ON THE FEATURES' CONTRIBUTION

After demonstrating in previous Sections the influence of FN method selection on the model's performance, next, as detailed in Section V-E1, an analysis of the features' contribution values estimated from the differently normalised datasets is conducted based on the traditional and the proposed adapted Garson's method. In addition, in order to demonstrate the superiority of the adapted Garson's method to truly infer the real features' contribution to the model, the FS strategy described in Section V-E2 is applied.

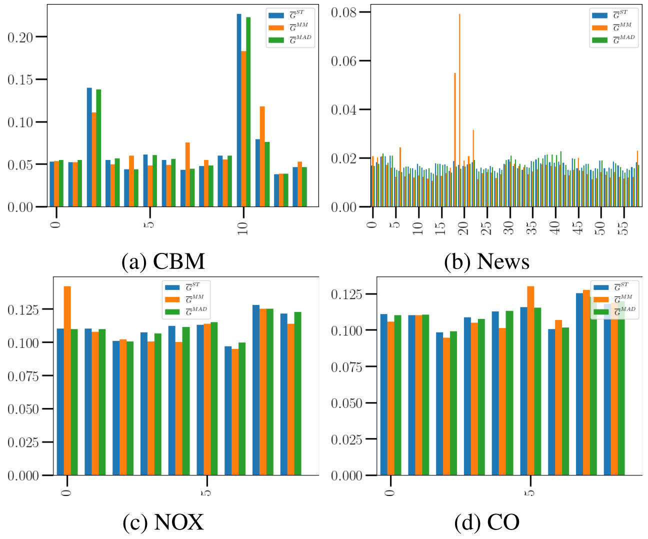
1) MEAN FEATURES' CONTRIBUTION

This Section analyses the dissimilarities between the features' contribution to the models trained with different FN methods, and the differences in the contribution values estimated with the traditional Garson's method and the proposed adapted one. As described in Section V-E, this inspection is conducted over the mean contribution values $\overline{G}$ and $\overline{\hat{G}}$ estimated from all the initialisation.

a: TRADITIONAL GARSON's METHOD

Fig. 5 depicts the values of $\overline{G}^*$ for $* \in$ *Norm*. In addition, Table 7a collects for each dataset, the difference between the highest and the lowest features' contribution values, and the std of each $\overline{G}^*$ are collected in Table 7b. $\tau(\overline{G}^*, \overline{G}^+)$ and $E_d(\overline{G}^*, \overline{G}^+)$ for $* \neq + \in$ *Norm* are shown in Tables 7c and 7d, respectively.

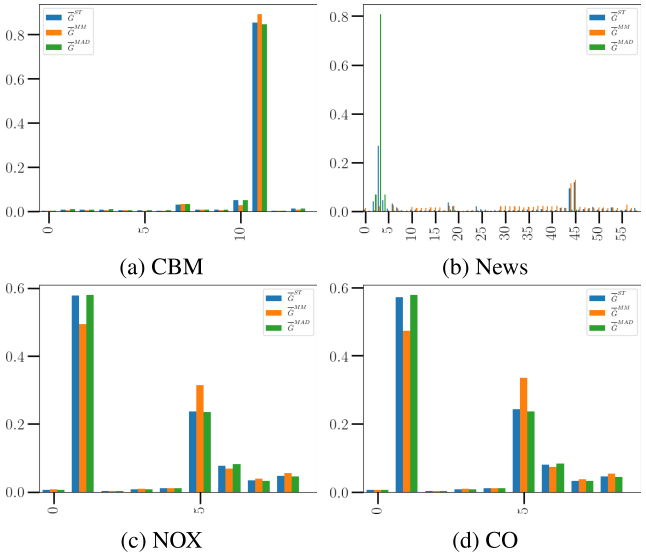
In Figs. 5a-5d it is observed that $\overline{G}^*$ values considerably varies depending on the FN method. In fact, Table 7a shows that the differences between the most extreme values are greater than 82% for the CBM dataset normalised with ST or MAD methods and for the News dataset normalised with MM. Contrary, in the other cases, the features with the lowest influence in the network present at least 60% the contribution value of the most influencing one. So, in these cases, the features' contribution to the network is more uniform than the observed in the former datasets. Finally, regarding the features' influence ranking, since 9 out of 12 Kendall's τ values are lower than 0.72 in Table 7c, it can be concluded that the selection of the FN method considerably alters the network's weight rank. The only cases with $\tau \geq 0.944$ are obtained when comparing $\overline{G}^{ST}$ and $\overline{G}^{MAD}$ for CBM, NOX and CO datasets. These results may be justified by the low difference between $\hat{w}^{ST}$ and $\hat{w}^{MAD}$ estimated for CBM, NOX and CO datasets observed in Figs. 4a and 4b, respectively.



(a) CBM (b) News

(c) NOX (d) CO

FIGURE 5. $\overline{G}^*$.



(a) CBM (b) News

(c) NOX (d) CO

FIGURE 6. $\overline{\hat{G}}^*$.

TABLE 7. Similitude analysis between the features' contribution estimated with the traditional Garson's method.

dataset	ST	MM	MAD
CBM	83.366	79.050	82.704
News	34.5	86.7	40.719
NOX	24.488	33.317	20.469
CO	21.605	27.399	19.233

(a) $max(\overline{G}^*) - min(\overline{G}^*)$.

dataset	ST	MM	MAD
CBM	0.051	0.04	0.05
News	0.002	0.01	0.002
NOX	0.01	0.015	0.009
CO	0.008	0.012	0.008

(b) $std(\overline{G}^*)$.

dataset	ST vs MM	ST vs MAD	MM vs MAD
CBM	0.187	0.956	0.231
News	0.385	0.716	0.405
NOX	0.444	1	0.444
CO	0.5	0.944	0.556

(c) Kendall's $\tau(\overline{G}^*, \overline{G}^+)$.

dataset	ST vs MM	ST vs MAD	MM vs MAD
CBM	0.077	0.006	0.075
News	0.078	0.01	0.08
NOX	0.036	0.005	0.037
CO	0.021	0.004	0.022

(d) $E_d(\overline{G}^*, \overline{G}^+)$.

TABLE 8. Similitude analysis between the features' contribution estimated with the adapted Garson's method.

dataset	ST	MM	MAD
CBM	99.789	99.863	99.759
News	99.749	98.711	99.994
NOX	99.457	99.318	99.492
CO	99.473	99.390	99.503

(a) $max(\overline{\hat{G}}^*) - min(\overline{\hat{G}}^*)$.

dataset	ST	MM	MAD
CBM	0.225	0.237	0.223
News	0.039	0.021	0.106
NOX	0.19	0.173	0.191
CO	0.189	0.171	0.19

(b) $std(\overline{\hat{G}}^*)$.

dataset	ST vs MM	ST vs MAD	MM vs MAD
CBM	0.824	0.978	0.846
News	0.204	0.799	0.092
NOX	1	1	1
CO	1	1	1

(c) Kendall's $\tau(\overline{\hat{G}}^*, \overline{\hat{G}}^+)$.

dataset	ST vs MM	ST vs MAD	MM vs MAD
CBM	0.046	0.008	0.053
News	0.268	0.563	0.818
NOX	0.114	0.006	0.118
CO	0.137	0.009	0.146

(d) $E_d(\overline{\hat{G}}^*, \overline{\hat{G}}^+)$.

b: PROPOSED ADAPTED GARSON's METHOD

In the following, the same analysis is performed over $\overline{\hat{G}}^*$ for $* \in Norm$.

Fig. 6 illustrates significant differences between $max\{\overline{\hat{G}}^*\}$ and $min\{\overline{\hat{G}}^*\}$ values. In fact, as Table 8a shows, the proportional differences between the most extreme contribution values are higher than 99%. This means that, in comparison with the most influencing feature, the feature with the lowest contribution value affects less than 1% the network's calculations. Regarding the Kendall's τ correlation coefficients collected in Table 6a, 5 out of 12 values are lower than 0.85, which means that, in those cases, the rank of $\overline{\hat{G}}^*$ varies depending on the FN method. In contrast, for NOX and CO datasets, the features' contribution ranks are the same independently from the normalisation method. This is coherent with the results observed in Table 6a, wherein $\hat{w}^*$ presented the same rank for $* \in Norm$.

c: COMPARISON BETWEEN THE TRADITIONAL AND THE PROPOSED ADAPTED GARSON's METHOD

Significant differences derived from the inclusion of dispersion factors in the features' relevance calculation are clear when examining Figs. 5 and 6. From the traditional Garson's method, the features present more uniformly distributed contribution values that the calculated ones with the proposed approach. In fact, all the std values from $\overline{G}^*$ are lower than 0.052 (Tables 7b); while, in Table 8b, the std values of $\overline{\hat{G}}^*$ are higher than 0.1 in most of the cases. Besides, in terms of Kendall's $\tau(\overline{G}^*, \overline{\hat{G}}^*)$ the importance rankings obtained with the traditional or the proposed Garson's methods are extremely different, as Table 9 illustrates.

In the following the results from Sections VI-C1.a and VI-C1.b are compared with those from Section VI-A.

Regarding the features' contribution values estimated with the proposed adapted Garson's method, the high τ and low E_d values for CBM dataset from Tables 8c and 8d may explain the mean MAE and RMSE differences close to 0 in Tables 3 and 4. In NOX, where the features' relevance rankings do not vary, the low $E_d(\overline{\hat{G}}^{ST}, \overline{\hat{G}}^{MAD})$ agree with the low mean MAE and RMSE differences from Table 3c for the corresponding case, while $E_d(\overline{\hat{G}}^{MM}, \overline{\hat{G}}^{ST})$= 0.114 and $E_d(\overline{\hat{G}}^{MM}, \overline{\hat{G}}^{MAD})$= 0.118 reflect the increment in the mean errors when comparing the resulting outputs. The same rationale is applied to the results obtained for

CO dataset. In contrast, for the mentioned cases, with the traditional Garson's method, the rank dissimilarities collected in Table 7c, and $E_d(\overline{G}^*, \overline{G}^+) < 0.1$ from Table 7d does not seem enough to explain the performance differences from Tables 3 and 4.

TABLE 9. $\tau(\overline{G}^*, \hat{\overline{G}}^*)$.

	ST	MM	MAD
CBM	0.165	0.495	0.187
News	0.161	0.316	0.003
NOX	0.222	−0.111	0.222
CO	0.222	0.5	0.278

Furthermore, contrary to the observed in Table 3b for News dataset, according to Tables 7c and 7d, the lowest mean MAE and RMSE errors would be expected from the dataset normalised with ST and MAD. However, the calculated errors in Table 3b agree with the trade-off between τ coefficients and Euclidean distance values from Tables 8c and 8d derived from the proposed adapted Garson's method.

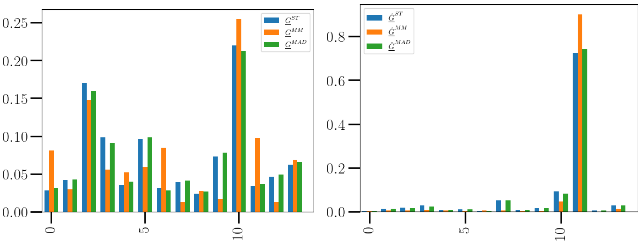
Then, it can be concluded that different features' contribution values are derived from the FN method selection and that higher correspondence exists between the results from Sections VI-A and VI-B respect to the features' contribution values estimated with the proposed adapted Garson's method compared to the observed with the traditional one.

2) FEATURE SELECTION BASED ON FEATURES' CONTRIBUTION

This Section applies the FS strategy described in Section V-E2 to demonstrate the superiority of the adapted Garson's method for estimating the true features' contribution to the model. For doing so, since multiple initialisations have been employed to train the models, for each $* \in Norm$, the model that reaches the lowest RMSE value is selected. Then, from such model, the features' contribution values computed with the traditional $\underline{G}^*$ and the proposed adapted Garson's method $\hat{\underline{G}}^*$ are employed.

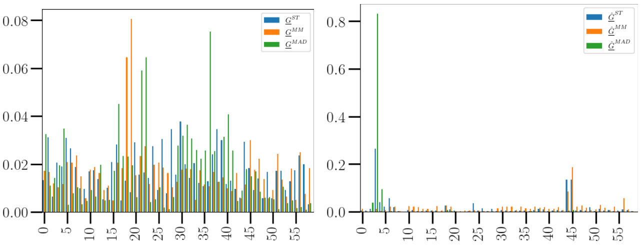
Figs. 7 to 10 depict the $\underline{G}^*$ and $\hat{\underline{G}}^*$ values estimated for each $* \in Norm$ and each dataset, respectively.

In Figs. 7b, 8b, 9b and 10b it is observed that the feature with lowest influence presents a contribution value lower than 1% the value of the highest contribution. Thus, respect to the most influencing one, the contribution to the model of at least one feature is insignificant. In fact, according to Figs. 7b and 8b, $\hat{\underline{G}}_j^* < 2 \cdot \max\{\hat{\underline{G}}_j^*\}$ for most of the features. In contrast, the features' contribution values estimated with the traditional Garson's method do not show as high disparity between the highest and lowest features contribution values. Besides, Table 10 collects the Kendall's τ coefficients from comparing the features' contribution rank estimated for $* \neq + \in Norm$ estimated with the traditional and the proposed Garson's method.



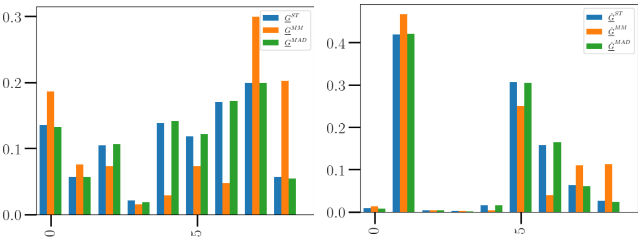
(a) $\underline{G}^*$ (b) $\hat{\underline{G}}^*$

FIGURE 7. CBM.



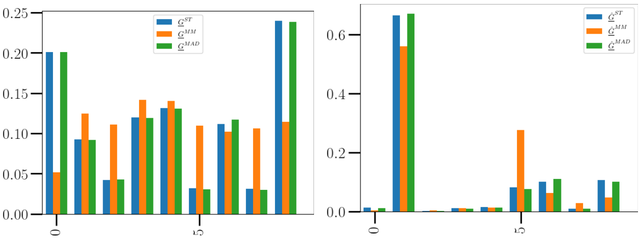
(a) $\underline{G}^*$ (b) $\hat{\underline{G}}^*$

FIGURE 8. News.



(a) $\underline{G}^*$ (b) $\hat{\underline{G}}^*$

FIGURE 9. NOX.



(a) $\underline{G}^*$ (b) $\hat{\underline{G}}^*$

FIGURE 10. CO.

When comparing the results from Tables 10a and 10b it is observed that the features' contribution rank significantly varies depending on $*$ according to the traditional Garson's method. In contrast, highest $\tau(\hat{\underline{G}}^*, \hat{\underline{G}}^+)$ values are obtained when comparing the features' rank estimated with the proposed approach.

Aiming at contrasting the features' contribution rank similarity estimated by the traditional and the proposed Garson's methods, Kendall's $\tau(\underline{G}^*, \hat{\underline{G}}^*)$ correlation coefficients are depicted in Table 11.

TABLE 10. For $* \neq + \in$ *Norm*, similarity between the features' contribution rank estimated by the traditional or the proposed Garson's method.

dataset	ST vs MM	ST vs MAD	MM vs MAD
CBM	0.165	0.956	0.165
News	0.065	0.084	0.056
NOX	0.167	0.944	0.111
CO	0.167	1	0.167

(a) Kendall's $\tau(\underline{G}^*, \underline{G}^+)$.

dataset	ST vs MM	ST vs MAD	MM vs MAD
CBM	0.56	0.978	0.582
News	0.244	0.341	0.218
NOX	0.778	1	0.778
CO	0.556	0.889	0.667

(b) Kendall's $\tau(\hat{\underline{G}}^*, \hat{\underline{G}}^+)$.

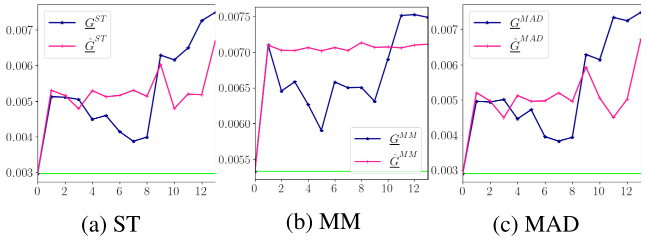
As Table 11 shows, since the τ values are lower than 0.5, there are significant differences in the feature's influence rankings when comparing both feature relevance analysis methods. In order to demonstrate the superiority of the proposed adapted Garson's method for the estimation of the real features' contribution values, the FS strategy (Algorithm 1) is applied.

FS is the strategy of removing disturbing or non-contributing features to improve the model's performance and reduce the computational cost and the memory requirements. As explained in Section IV-B and proven in Section VI, this work states and demonstrates the influence of the FN method selection in the model's performance and in the features' contribution to the model. Thus, in this Section the features removal is conducted as described in Section V-E2 based on $\underline{G}^*$ and $\hat{\underline{G}}^*$ for $* \in$ *Norm*. Every time a feature is discarded, the model is retrained, and the RMSE between the estimated output and the real one is calculated. This experiment aims to compare the validity of the adapted Garson's method, against the traditional Garson's method, for estimating the real features' contribution. For each dataset and each $*$, the random initialisation that reaches the lowest RMSE when employing the whole dataset is utilised. Note that given a dataset, the lowest RMSE value is obtained with different random initialisations for the different FN methods.

TABLE 11. Kendall's τ correlation between $\tau(\underline{G}^*$ and $\hat{\underline{G}}^*)$ for each $* \in$ *Norm*.

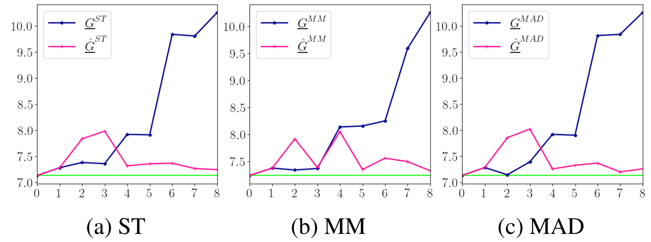
	ST	MM	MAD
CBM	0.363	0.495	0.297
News	0.37	0.404	0.433
NOX	0.167	0.444	0.222
CO	0.222	0.056	0.111

Figs. 11 to 13 depict for each dataset and each FN method the RMSE value obtained for each iteration of the FS strategy. The X-axis refers to the number of features removed at each stage of the procedure. Thus, 0 refers to the employment of the whole dataset. The Y-axis collects the RMSE value between the real and the estimated output for the training set. The blue stars depict the results obtained when the features are discarded according to $\underline{G}^*$; and the pink vertical lines, the RMSE value resulting from the feature selection strategy based on $\hat{\underline{G}}^*$. The horizontal green line represents the RMSE value reached with the complete dataset. Note that the features are removed one by one, and since the rank similarity between the contributions estimated by the traditional and the adapted Garson's methods differs, the removed feature may not coincide at each stage of the algorithm.



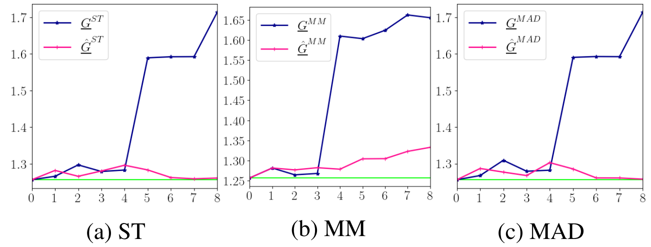
(a) ST (b) MM (c) MAD

FIGURE 11. CBM dataset.



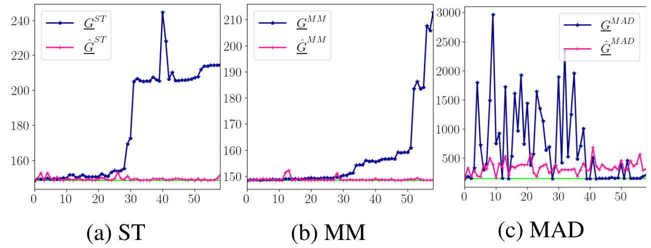
(a) ST (b) MM (c) MAD

FIGURE 12. NOX dataset.



(a) ST (b) MM (c) MAD

FIGURE 13. CO dataset.



(a) ST (b) MM (c) MAD

FIGURE 14. News dataset.

When comparing CBM, NOX and CO datasets it is observed that the RMSE values resultant from the features removal based on $\underline{G}^{ST}$ and $\underline{G}^{MAD}$ and based on $\hat{\underline{G}}^{ST}$ and $\hat{\underline{G}}^{MAD}$ are approximately the same. This was expected from the results of Tables 10a and 10b. Nevertheless, in these cases, especially for NOX and CO datasets, the RMSE values obtained from the FS based on the adapted Garson's method are closer to the performance reached with the whole dataset, especially when increasing the number of removed features. In fact, as observed in Figs. 7a, 9a and 10a, and in Table 10b, there are significant differences between the contribution value estimated for the most influencing features and the resting ones. Consequently, in Figs. 12a to 12c and 13a to 13c it is observed that the RMSE obtained with all the features and the RMSE obtained when utilising uniquely the most influencing features is almost the same. Furthermore, in News dataset, in Fig. 4c significant differences between $\hat{w}^*$ for $* \in Norm$ were observed. Consequently, in Figs. 14a to 14c by comparing the results from the FS strategy according to the traditional or the adapted Garson's methods, it is clear that each feature contribution in the model differs depending on the normalisation method employed to transform News dataset. Besides, in Figs. 14a and 14b it is observed that the RMSE error estimated at each stage of the FS strategy based on the $\underline{\hat{G}}^*$ remains closer to the RMSE obtained with all the features than the RMSE resulting from FS according to Garson's traditional method.

All in all, it is demonstrated that the dispersion factors inclusion in the features' contribution calculation significantly improves the estimation of the real features' influence on the model, as observed through the FS strategy.

VII. DISCUSSION

As stated and demonstrated in this work, the FN method selection significantly affects the ANN-based model's performance and the inclusion of dispersion factors when estimating the features' contribution improves the understanding of the features' influence on the model.

The former point emphasises the influence of the FN method selection; however, it remains open the question about which FN to employ to transform a given dataset in order to reach the best model's performance; or even if it is recommendable the application of FN or discard the magnitude of the features by removing the 10^{n_j} factors from (4). As stated in Sections III and IV-B, FN imposes a dispersion weight to compress or expand the features. Thus, FN can be viewed as a Feature Weighting method that estimates the features' weights in an unsupervised manner since the dispersion factors are calculated based on the features' statistical characteristics. A weight that does not correspond to the real relative importance of a given feature can result in a performance loss. In fact, in Table 4b it is observed that for the test set, lower mean RMSE and higher R^2 scores are obtained from the raw dataset with the magnitude factors removed than from any normalised dataset. Thus, further research about the suitability of the FN method selection would be interesting given the properties of a given dataset. Moreover, since this work demonstrates the influence of the FN on the network, it is evident that other preprocessing techniques may also condition the model's performance. Hence, the impact of supervised FW preprocessing methods to improve the model's performance should be investigated. Furthermore, a conjoint comparison between the supervised weights calculated with a given FW method and their similitude with the dispersion factors estimated with different FN may guide the selection of a given normalisation method to preprocess the input data.

In addition, this work analyses the weight matrix analysis-based methods to understand the features' contribution to the model. However, it would be interesting to extend the analysis to other explainability analysis approaches. The presented results are obtained from networks with the identity activation function in the hidden and output layers. Then, further studies for network's with different activation functions are needed.

Another interesting research topic until the date in the ANN branch is the search of the optimal weights initialisation to maintain the fair initial features' contribution to the solution search space. However, in the same way that FN influences the features' contribution to the model's performance, it may be suspected that it may also condition the suitability of the initial weight configuration for a fair weights adjustment. As aforementioned, the lowest RMSE values reached by each normalised dataset are obtained with different random initialisations. Moreover, note that despite the significant differences in terms of mean MAE, RMSE and R^2 based on $*$; the minimum estimated errors (reached with different initialisations) in Table 4 are almost the same for each normalised dataset. Further studies about the network initialisation based on the conjoint influence of the dispersion factors and the initial weight matrix may be of great interest, which may result in a new weight initialisation strategy.

VIII. CONCLUSION

Due to the high ability of ANN to model complex systems, these algorithms are being widely employed to solve complex problems. Simultaneously, because of the lack of explainability of the ANN, state-of-the-art focuses on bringing some understanding about the network functioning. In this field, several works aim at analysing the features' contribution to the model via weight matrices study. However, in such works, the preprocessing phase is not considered when estimating the features' contribution. This work has been theoretically proven and later experimentally validated that the dispersion factors employed to transform the input features' influence the final features' contribution to the model and the model's performance. In fact, as shown in this work, the presented proportional dispersion weights are explanatory factors of the similarity between the performance obtained by models trained with different FN methods. Then, as a conclusion of this work, it is recommended to include information about the dispersion factors to analyse the features' real contribution. In this line, this work proposes adapted Garson's and Yoon's methods that include features' dispersion factors for a more precise estimation of the features' influence on the model. Besides, a feature selection strategy is employed to analyse in terms of RMSE variations the effect of removing features according to Garson's method or the proposed adapted Garson's method. These experiments demonstrate that the RMSE results obtained when removing features according to the adapted Garson's method match the conclusions obtained from the features' contribution values. Then, the knowledge extracted from this proposal improves the understanding of the features' contribution to the model and enhances the feature selection strategy, which is fundamental in real use cases to model the problem at hand reliably.

Future work will focus on considering the conjoint comparison between the features' weights derived from supervised

FW and their similitude with the dispersion factors to guide the optimal FN method selection. Besides, the impact of FW as preprocessing technique for performance improvement and the influence of preprocessing techniques on ANNs with different activation functions will be considered in future works. Moreover, new network's initialisation approaches based on the conjoint influence of the preprocessing factors and the initial weight matrix may be an exciting future research topic.

REFERENCES

[1] M. H. Hassoun, *Fundamentals of Artificial Neural Networks*. Cambridge, MA, USA: MIT Press, 1995.

[2] A. K. Jain, J. Mao, and K. M. Mohiuddin, "Artificial neural networks: A tutorial," *Computer*, vol. 29, no. 3, pp. 31–44, Mar. 1996.

[3] O. I. Abiodun, M. U. Kiru, A. Jantan, A. E. Omolara, K. V. Dada, A. M. Umar, O. U. Linus, H. Arshad, A. A. Kazaure, and U. Gana, "Comprehensive review of artificial neural network applications to pattern recognition," *IEEE Access*, vol. 7, pp. 158820–158846, 2019.

[4] C. M. Agu, M. C. Menkiti, E. B. Ekwe, and A. C. Agulanna, "Modeling and optimization of Terminalia catappa L. Kernel oil extraction using response surface methodology and artificial neural network," *Artif. Intell. Agricult.*, vol. 4, pp. 1–11, Jan. 2020.

[5] M. Jalanko, Y. Sanchez, V. Mahalec, and P. Mhaskar, "Adaptive system identification of industrial ethylene splitter: A comparison of subspace identification and artificial neural networks," *Comput. Chem. Eng.*, vol. 147, Apr. 2021, Art. no. 107240.

[6] J. Lee, Y. C. Lee, and J. T. Kim, "Migration from the traditional to the smart factory in the die-casting industry: Novel process data acquisition and fault detection based on artificial neural network," *J. Mater. Process. Technol.*, vol. 290, Apr. 2021, Art. no. 116972.

[7] B. Li, C. Li, J. Huang, and C. Li, "Application of artificial neural network for prediction of key indexes of corn industrial drying by considering the ambient conditions," *Appl. Sci.*, vol. 10, no. 16, p. 5659, Aug. 2020.

[8] Y. Park, M. Choi, K. Kim, X. Li, C. Jung, S. Na, and G. Choi, "Prediction of operating characteristics for industrial gas turbine combustor using an optimized artificial neural network," *Energy*, vol. 213, Dec. 2020, Art. no. 118769.

[9] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.

[10] A. Fischer, "How to determine the unique contributions of input-variables to the nonlinear regression function of a multilayer perceptron," *Ecol. Model.*, vols. 309–310, pp. 60–63, Aug. 2015.

[11] J. D. Olden and D. A. Jackson, "Illuminating the 'black box': A randomization approach for understanding variable contributions in artificial neural networks," *Ecol. Model.*, vol. 154, nos. 1–2, pp. 135–150, 2002.

[12] C. R. D. Sá, "Variance-based feature importance in neural networks," in *Proc. Int. Conf. Discovery Sci.* Cham, Switzerland: Springer, 2019, pp. 306–315.

[13] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020.

[14] A. Eck, L. M. Zintgraf, E. F. J. de Groot, T. G. J. de Meij, T. S. Cohen, P. H. M. Savelkoul, M. Welling, and A. E. Budding, "Interpretation of microbiota-based diagnostics by explaining individual classifier decisions," *BMC Bioinf.*, vol. 18, no. 1, pp. 1–13, Dec. 2017.

[15] K. Amarasinghe and M. Manic, "Explaining what a neural network has learned: Toward transparent classification," in *Proc. IEEE Int. Conf. Fuzzy Syst. (FUZZ-IEEE)*, Jun. 2019, pp. 1–6.

[16] D. Janzing, L. Minorics, and P. Blöbaum, "Feature relevance quantification in explainable AI: A causal problem," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2020, pp. 2907–2916.

[17] J. Jiménez-Luna, F. Grisoni, and G. Schneider, "Drug discovery with explainable artificial intelligence," *Nature Mach. Intell.*, vol. 2, no. 10, pp. 573–584, Oct. 2020.

[18] J. Wang, J. Wiens, and S. Lundberg, "Shapley flow: A graph-based approach to interpreting model predictions," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2021, pp. 721–729.

[19] B. Venkatesh and J. Anuradha, "A review of feature selection and its methods," *Cybern. Inf. Technol.*, vol. 19, no. 1, pp. 3–26, Mar. 2019.

[20] V. Bolón-Canedo and A. Alonso-Betanzos, "Ensembles for feature selection: A review and future trends," *Inf. Fusion*, vol. 52, pp. 1–12, Dec. 2019.

[21] Q. Al-Tashi, S. J. Abdulkadir, H. M. Rais, S. Mirjalili, and H. Alhussian, "Approaches to multi-objective feature selection: A systematic literature review," *IEEE Access*, vol. 8, pp. 125076–125096, 2020.

[22] C. M. Bishop, *Neural Networks for Pattern Recognition*. London, U.K.: Oxford Univ. Press, 1995.

[23] K. D. Pramanik, M. Mukhopadhyay, and S. Pal, "Big data classification: Applications and challenges," in *Artificial Intelligence and IoT: Smart Convergence for Eco-Friendly Topography*, vol. 85. Singapore: Springer, 2021, p. 53. [Online]. Available: https://www.springer.com/gp/book/9789813363991

[24] Y. Kong and T. Yu, "A graph-embedded deep feedforward network for disease outcome classification and feature selection using gene expression data," *Bioinformatics*, vol. 34, no. 21, pp. 3727–3737, Nov. 2018.

[25] O. Yucel, E. S. Aydin, and H. Sadikoglu, "Comparison of the different artificial neural networks in prediction of biomass gasification products," *Int. J. Energy Res.*, vol. 43, no. 11, pp. 5992–6003, Sep. 2019.

[26] H. Turabieh, M. Mafarja, and X. Li, "Iterated feature selection algorithms with layered recurrent neural network for software fault prediction," *Expert Syst. Appl.*, vol. 122, pp. 27–42, May 2019.

[27] F. Curreri, G. Fiumara, and M. G. Xibilia, "Input selection methods for soft sensor design: A survey," *Future Internet*, vol. 12, no. 6, p. 97, Jun. 2020.

[28] B. Jeong and H. Cho, "Feature selection techniques and comparative studies for large-scale manufacturing processes," *Int. J. Adv. Manuf. Technol.*, vol. 28, nos. 9–10, pp. 1006–1011, Jul. 2006.

[29] N. L. da Costa, M. D. de Lima, and R. Barbosa, "Evaluation of feature selection methods based on artificial neural network weights," *Expert Syst. Appl.*, vol. 168, Apr. 2021, Art. no. 114312.

[30] T. A. Folorunso, A. M. Aibinu, J. G. Kolo, S. O. Sadiku, and A. M. Orire, "Effects of data normalization on water quality model in a recirculatory aquaculture system using artificial neural network," *i-Manager's J. Pattern Recognit.*, vol. 5, no. 3, p. 21, 2018.

[31] A. Gökhan, C. O. Güzeller, and M. T. Eser, "The effect of the normalization method used in different sample sizes on the success of artificial neural network model," *Int. J. Assessment Tools Educ.*, vol. 6, no. 2, pp. 170–192, Apr. 2019.

[32] B. K. Singh, K. Verma, and A. Thoke, "Investigations on impact of feature normalization techniques on classifier's performance in breast tumor classification," *Int. J. Comput. Appl.*, vol. 116, no. 19, pp. 11–15, Apr. 2015.

[33] X. A. Larriva-Novo, M. Vega-Barbas, V. A. Villagrá, and M. S. Rodrigo, "Evaluation of cybersecurity data set characteristics for their applicability to neural networks algorithms detecting cybersecurity anomalies," *IEEE Access*, vol. 8, pp. 9005–9014, 2020.

[34] C. E. Choong, S. Ibrahim, and A. El-Shafie, "Artificial neural network (ANN) model development for predicting just suspension speed in solid-liquid mixing system," *Flow Meas. Instrum.*, vol. 71, Mar. 2020, Art. no. 101689.

[35] N. Khare, P. Devan, C. Chowdhary, S. Bhattacharya, G. Singh, S. Singh, and B. Yoon, "SMO-DNN: Spider monkey optimization and deep neural network hybrid classifier model for intrusion detection," *Electronics*, vol. 9, no. 4, p. 692, Apr. 2020.

[36] R. Ramani, K. V. Devi, and K. R. Soundar, "MapReduce-based big data framework using modified artificial neural network classifier for diabetic chronic disease prediction," *Soft Comput.*, vol. 24, no. 21, pp. 16335–16345, Nov. 2020.

[37] W. Zhang, Q. M. J. Wu, Y. Yang, and T. Akilan, "Multimodel feature reinforcement framework using Moore–Penrose inverse for big data analysis," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Oct. 6, 2020, doi: 10.1109/TNNLS.2020.3026621.

[38] Z. Shi, W. Zheng, and W. Yin, "Improving the reliability of the prediction of terrestrial water storage in Yunnan using the artificial neural network selective joint prediction model," *IEEE Access*, vol. 9, pp. 31865–31879, 2021.

[39] D. Dua and C. Graff, "UCI machine learning repository," School Inf. Comput. Sci., Univ. California, Irvine, CA, USA, 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[40] S. García, J. Luengo, and F. Herrera, *Data Preprocessing in Data Mining*, vol. 72. Cham, Switzerland: Springer, 2015.

[41] G. D. Garson, "Interpreting neural-network connection weights," *AI Expert*, vol. 6, no. 4, pp. 46–51, Apr. 1991, doi: 10.5555/129449.129452.

[42] Y. Yoon, G. Swales, Jr., and T. M. Margavio, "A comparison of discriminant analysis versus artificial neural networks," *J. Oper. Res. Soc.*, vol. 44, no. 1, pp. 51–60, Jan. 1993.

[43] A. Coraddu, L. Oneto, A. Ghio, S. Savio, D. Anguita, and A. Figari, "Machine learning approaches for improving condition-based maintenance of naval propulsion plants," *Proc. Inst. Mech. Eng., M, J. Eng. Maritime Environ.*, vol. 230, no. 1, pp. 136–153, 2016.

[44] H. Kaya, P. Tüfekci, and E. Uzun, "Predicting CO and NO$_x$ emissions from gas turbines: Novel data and abenchmark PEMS," *TURKISH J. Electr. Eng. Comput. Sci.*, vol. 27, no. 6, pp. 4783–4796, Nov. 2019.

[45] K. Fernandes, P. Vinagre, and P. Cortez, "A proactive intelligent decision support system for predicting the popularity of online news," in *Proc. Portuguese Conf. Artif. Intell.* Cham, Switzerland: Springer, 2015, pp. 535–546.

[46] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, 2010, pp. 249–256.

[47] L. Puka, *Kendall's Tau*. Berlin, Germany: Springer, 2011, pp. 713–715.

[48] R. Woolson, "Wilcoxon signed-rank test," in *Wiley Encyclopedia of Clinical Trials*. Hoboken, NJ, USA: Wiley, 2007, pp. 1–3. [Online]. Available: https://www.springer.com/gp/book/9789813363991

IRATXE NIÑO-ADAN was born in Bilbao, Basque Country, in 1990. She received the bachelor's degree in mathematics and the master's degree in mathematical modeling and research, statistics, and computing from the University of the Basque Country (UPV/EHU), in 2015 and 2017, respectively. She is currently pursuing the Ph.D. degree with Tecnalia Research and Innovation. Her research interests include developing advanced techniques with regards to data analytics and Industry 4.0.

EVA PORTILLO received the Ph.D. degree in engineering, in 2007. She is currently an Associate Professor with the Department of Automatic Control and Systems Engineering, University of the Basque Country (UPV/EHU). She has several awards at national conferences. In 2016, she was a Visiting Professor with the Knowledge Engineering and Discovery Research Institute (KEDRI), Auckland University of Technology, Auckland, New Zealand. From 2017 to 2019, she has been the Vice-Dean of the master's and Doctoral School of the University of the Basque Country, where she is currently an Academic Secretary of the Doctoral School. She received the Prize for Outstanding Ph.D. thesis awarded by UPV/EHU, in 2010.

ITZIAR LANDA-TORRES received the Ph.D. degree in telecommunication engineering from the University of Deusto and the Ph.D. degree in information technology from the University of Alcalá de Henares (UAH). She has been a Researcher at Tecnalia Research and Innovation, working in artificial intelligence, data-mining, pattern analysis, neural networks, clustering, and grouping problems related to different fields of knowledge. During her ten-year-long research career, she has coauthored more than 25 scientific publications in various international journals *Applied Soft Computing*, *Engineering Applications of Artificial Intelligence*, and *Expert Systems with Applications*. She is currently the Innovation Project Manager at Petronor Innovación S.L. (Repsol).

DIANA MANJARRES received the Ph.D. degree in telecommunication engineering from the University of the Basque Country (UPV/EHU) and the Ph.D. degree in information technology from the University of Alcalá (UAH). She is currently a Scientific Researcher at Tecnalia Research and Innovation, working in artificial intelligence and optimization algorithms. During her ten-year-long research career, she has coauthored more than 25 scientific publications in various international journals and conferences. Her research interests include heuristic techniques for NP-hard optimization problems, multi-objective optimization, data-mining, pattern analysis, neural networks, clustering, and grouping problems related to different fields of knowledge.