Surrogate-Assisted and Filter-Based Multiobjective Evolutionary Feature Selection for Deep Learning

Raquel Espinosa, Fernando Jiménez, and José Palma, *Senior Member, IEEE*

***Abstract*— Feature selection (FS) for deep learning prediction models is a difficult topic for researchers to tackle. Most of the approaches proposed in the literature consist of embedded methods through the use of hidden layers added to the neural network architecture that modify the weights of the units associated with each input attribute so that the worst attributes have less weight in the learning process. Other approaches used for deep learning are filter methods, which are independent of the learning algorithm, which can limit the precision of the prediction model. Wrapper methods are impractical with deep learning due to their high computational cost. In this article, we propose new attribute subset evaluation FS methods for deep learning of the wrapper, filter and wrapper-filter hybrid types, where multiobjective and many-objective evolutionary algorithms are used as search strategies. A novel surrogate-assisted approach is used to reduce the high computational cost of the wrapper-type objective function, while the filter-type objective functions are based on correlation and an adaptation of the reliefF algorithm. The proposed techniques have been applied in a time series forecasting problem of air quality in the Spanish south-east and an indoor temperature forecasting problem in a domotic house, with promising results compared to other FS techniques used in the literature.**

***Index Terms*— Air quality, deep learning, feature selection (FS), indoor temperature, multiobjective evolutionary algorithms (MOEAs), surrogate-assisted, time series forecasting.**

I. Introduction

The increase in the amount and complexity of available data leads to an increase in the dimensionality of the information. *Feature selection* (FS) [1] is a process that reduces the dimensionality of the input data, as well as the complexity of the models created from the input data. However, very few works focus on FS in deep learning. Most of them consist of embedded methods that integrate FS into the training process of deep learning models. Other FS methods widely used with deep learning are filter-type. They are fast methods, but independent of the learning algorithm, so they are good in general but not so accurate for a given learning algorithm. On the other hand, wrapper methods are very accurate but very expensive, especially when the learning algorithm is slow and dataset contains a large number of input attributes. For this reason, little research has been carried out. There are also wrapper-filter hybrids and ensembles for FS.

Therefore, we propose different multiobjective optimization models for FS both wrapper and filter type, and also hybrid wrapper-filter models. The proposed optimization models contain 2, 3, or 4 objectives, which are solved with *multiobjective evolutionary algorithms* (MOEAs) [2]. For the evaluation of the objective functions of the wrapper-type, the *root mean squared error* (RMSE) of a surrogate for a *long short-term memory* (LSTM) [3] artificial recurrent neural network is used. Filter-type objectives are based on correlation metrics and an adaptation of the reliefF method for evaluating subsets of attributes. All the proposed optimization models contain an objective for minimizing the number of attributes.

The proposed methods have been applied to a time series forecasting problem of air quality. In particular, to the prediction of the concentration of nitrogen dioxide (NO_2) in the town of La Aljorra, Region of Murcia, Spain. NO_2 is a noxious gas that is produced in combustion processes such as vehicle engines or some industries. In addition, NO_2 interacting with other chemicals in the air like water or oxygen can cause acid rain. Prolonged exposure to this chemical compound can lead to respiratory [4] and cardiovascular [5] diseases. Additionally, a problem of indoor temperature forecasting in a domotic house in Valencia (Spain) has also been analyzed to strengthen the conclusions.

The main contributions of this work are, in summary, the following.

1) We propose new multiobjective FS methods of wrapper, filter, and hybrid types based on LSTM recurrent neural networks, correlation, the reliefF algorithm and the minimization of the number of attributes.
2) We propose a novel surrogate-assisted MOEA to solve the proposed wrapper and hybrid FS methods.
3) We apply the proposed FS methods to the forecast of time series for air quality in the city of La Aljorra (Spain), and in a problem of indoor temperature forecasting in a domotic house in Valencia (Spain), using a future prediction horizon to h-step ahead.

Manuscript received 19 October 2021; revised 28 July 2022, 25 October 2022, and 23 December 2022; accepted 4 January 2023. This work was supported in part by the CONFAINCE Project under Grant PID2021-122194OB-I00, in part by the Spanish Ministry of Science and Innovation, in part by the Spanish Agency for Research, in part by the IMPACT-T2D Project under Grant PMP21/00092, and in part by the Spanish Health Institute Carlos III (ISCIII). *(Corresponding author: Raquel Espinosa.)*

Raquel Espinosa is with the International Doctorate School, University of Murcia, 30100 Murcia, Spain (e-mail: raquel.espinosa@um.es).

Fernando Jiménez and José Palma are with the Department of Information Engineering and Communications, University of Murcia, 30100 Murcia, Spain.

This article has supplementary material provided by the authors and color versions of one or more figures available at https://doi.org/10.1109/TNNLS.2023.3234629.

Digital Object Identifier 10.1109/TNNLS.2023.3234629

4) We propose a decision-making mechanism for choosing the non-dominated solution from the Pareto front, specific for time series forecasting problems.
5) We propose a methodology to compare multiobjective optimization models for FS, time series forecasting models, and multiobjective and many-objective evolutionary algorithms, based on non-parametric statistical tests on appropriate performance metrics and multicriteria aggregation functions at h-step ahead prediction horizons. Our models are compared with the models obtained using other FS techniques used in literature, such as linear regression-based wrapper FS methods, random forest-based wrapper FS methods, correlation-based feature ranking (FR) methods, reliefF, and embedded FS methods.
6) The proposed method reduces the run time of the FS process, allows dealing with high dimensional problems, and reduces the complexity of the prediction models. This reduction in computing time implies a reduction in the carbon footprint, helping to deal with climate change. From a social and economic point of view, our proposal is in line with the objectives of the *Paris Agreement* on climate change (https://www.un.org/en/climatechange/paris-agreement), the *Sustainable Development Goals* (https://sdgs.un.org/) and the *European Green Deal* (https://ec.europa.eu/info/strategy/priorities-2019-2024/european-green-deal). On the other hand, reducing the number of attributes in prediction models contributes to *explainable artificial intelligence* (XIA) [6], which is currently in great social demand and is part of the objectives of, among others, the *White Paper on Artificial Intelligence* of the European Commission (https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust) and the *Executive Order 13960 on Promoting the Use of Trustworthy AI in the Federal Government* of USA (https://www.federalregister.gov/documents/2020/12/08/2020-27065/promoting-the-use-of-trustworthy-artificial-intelligence-in-the-federal-government).

The rest of the article is organized as follows. Section II presents some background for FS, surrogate-assisted evolutionary computation, and air quality; Section III describes the materials and methods used in this work; Section IV presents the performed experiments and shows the results obtained; Section V analyses and discusses the results; Section VI draws conclusions and future work; finally, Appendixes A–C show abbreviations, deep learning architectures and tables that reinforce the results.

II. Background

Section II presents the background and describes some of the works that relate FS to deep learning and multiobjective metaheuristics, as well as work related to surrogate-assisted evolutionary algorithms and air quality time series forecasting.

A. Feature Selection

FS is the process of removing irrelevant and redundant attributes from a database [7]. The goal of FS is to find the best subset of attributes, reducing the cost and computational time required for model learning. Another interesting property is that it can improve the interpretability of the models by making them less complex and therefore easier to understand and interpret. A FS process can be either an attribute subset or an attribute evaluation process. The former searches for subsets of attributes using some search strategy. Their advantage is that they consider the interaction between attributes (multivariate methods), but the search requires a number of steps that is $O(2^n)$ (NP-hard problem) with n being the number of attributes. The latter evaluates attributes in order to assign them a ranking or importance, thus they are called FR methods [8]. These methods are usually univariate, although there are a few multivariate methods, such as *reliefF* [9]. FR techniques can also be treated as FS techniques if a subset with the best attributes in the ranking, or those above a certain threshold of importance, is selected. Zhang et al. [10] propose a $l_{0,2}-norm$ based FS method to select the top k features in various scenarios. The subset evaluation FS process has several stages. First, from the initial database, a subset of attributes is selected with a search strategy. Some of the most popular search strategies are *forward FS* [11], *recursive feature elimination* [12] or *multiobjective evolutionary search* [13]. The subsets generated by the search strategy are evaluated with some evaluation function. If after evaluation a stop criterion is satisfied, the process is terminated and moves to a validation phase. If the stop criterion is not satisfied, a new subset is generated again until the stop criterion is satisfied.

FS methods can typically be classified into three types: filter, wrapper, and embedded. Filter methods evaluate attributes through statistical functions such as *correlation* [14], *consistency* [15], *redundancy* [16], *information gain* [17], χ^2 [18], and so on. They are the fastest and simplest. Wrapper methods build machine learning models whose performance is evaluated using some metric, such as *accuracy* and *area under the receiver operating characteristic (ROC) curve* for classification tasks, or RMSE and *mean absolute error* (MAE) for regression. Therefore, the computational cost is usually high, although they are, in general, more accurate than filter FS methods. *Embedded* methods perform FS within the model training process. Examples of this can be *least absolute shrinkage and selection operator* (LASSO) [19], *C4.5* [20] or *classification and regression trees* (CARTs) [21]. *Ensembles* can also be used as an FS technique [22]. They are a combination of the results obtained by multiple techniques. Ensembles try to achieve better outcomes than would be expected using only one method. Finally, there are *hybrid methods* that generally combine filter and wrapper. The purpose of this is usually to improve the prediction error of the filters and reduce the computational cost of the wrapper.

FS in deep learning models is less common than in machine learning. However, in the literature, we can find several articles on FS with deep learning. Embedded methods tend to be the most common when performing FS with deep learning, as a layer can be added to the neural network to perform this process. For example, Borisov et al. [23] have developed a new input layer for *artificial neural networks* (ANNs), called *CancelOut*, for FS in deep learning models. Other examples of neural networks that integrate FS are *NeuralFS* [24], based on DNN, and CNN-FS [25], based on *convolutional neural networks* (CNNs) [26]. Wang et al. [27] combines CNN for feature extraction with χ^2 for FS, along with ensemble learning based on *extreme machine learning* (EML) and weighted voting. Attention-based methods are another effective method for handling features that have different importance in deep learning. Yuan et al. [28] propose a spatio-temporal attention-based LSTM network that can identify important input variables at each time step and adaptively discover hidden quality-related states at all time steps. Wang et al. [29] propose an integrated FS scheme based on a neural network with redundancy control and Group Lasso regularization to produce sparsity.

There are also modifications of embedded FS methods to consider discarded features such as the work of Zheng et al. [30], who propose an embedded framework for FS where an unselected feature classifier to find an optimal subset of attributes is added. In this line of research, there are other methods to retain more statistical and structural information about the features such as the case of Wu et al. [31] where an orthogonal least square regression model with feature weighting for FS is proposed.

FS has been successfully applied in environments where features are not previously known to the learner but are streams whose information is increasing over time, as is the case Zhou et al. [32] propose a streaming FS considering feature interaction.

B. Multiobjective Meta-Heuristics for FS Based Deep Learning

For the design of FS methods, the use of MOEAs is very popular. MOEAs simultaneously optimize several conflicting objective functions. This gives rise to the set of Pareto optimal solutions, i.e., solutions where one objective cannot be further improved without making another objective worse. The optimal solutions in a multiobjective optimization problem are called *non-dominated solutions*, and they form the *Pareto front*. Evolutionary algorithms are well suited for multiobjective optimization because they allow multiple non-dominated solutions to be captured in a single run of the algorithm, in addition to their ability to solve complex problems. MOEAs are called *many-objective evolutionary algorithms* [33] when solving optimization problems of more than three objectives.

Some works use multiobjective optimization algorithm-based deep learning to perform FS. For example, the one proposed by Al-Tashi et al. [34] who present a wrapper based on ANN for FS in classification, with the *multiobjective gray wolf optimizer* (BMOGW) as the search strategy. Within the hybrid methods with deep learning, we can find combinations of filters and wrappers. For financial prediction, Niu et al. [35] combines *reliefF* filter and a wrapper based on EML, and BMOGW combined with *cuckoo search* to perform FS. Although multiobjective optimization is the most common, there are also studies focusing on many-objective optimization. Recently, Shu et al. [36] have proposed a five-objective optimization model with *non-dominated sorting genetic algorithm III* (NSGA-III) [37], of which four are filters and the remaining one is a wrapper based on EML.

C. Surrogate-Assisted Evolutionary Algorithms

Surrogate-assisted evolutionary computation attempt to approximate the fitness function in evolutionary algorithms through a more computationally efficient model [38]. This technique is especially useful when dealing with high-dimensional problems and has proven to be effective in multiobjective optimization problems [39], [40].

An example of a surrogate-assisted method with deep learning and multiobjective meta-heuristics used to perform FS is shown in the work of Jiang et al. [41]. They propose an ensemble method based on a two-layer surrogate-assisted mechanism called *multisurrogate-assisted dual-layer ensemble FS* (MDEFS). To select the most relevant features in large datasets, a subset with the most representative samples is selected with three strategies based on surrogate-assisted models: k-means, *density-based spatial clustering of applications with noise* and *random sub-sampling*. Then, a *particle swarm*-based algorithm is used to select the best attributes.

D. Air Quality

Air quality forecasting with multiobjective evolutionary computation combined with deep learning is a field that is still under explored. However, there are already some papers that address this issue. Most of them are focused on the predictions of $PM_{2.5}$ in different areas of China, such as Liu et al. [42], that introduce a multiresolution ensemble model based on *wavelet packet decomposition*, *bidirectional* LSTM and *non-dominated sorting genetic algorithm II* (NSGA-II) [43], which tries to minimize bias and variance, to obtain deterministic forecasts. Zhang et al. [44] propose an ensemble model based on *multiobjective ensemble pruning* with *multipopulation* NSGA-II and MLP, CNN, and LSTM for stable $PM_{2.5}$ time series forecasting. Wang et al. [45] developed a hybrid system based on two-step FS with correlation, *reliefF* and a modified *quantum fuzzy neural network* for air quality forecasting. The relevance of the selected attributes is determined through a multiobjective chaotic map-based algorithm called *multiobjective chaotic bonobo optimizer*. Du et al. [46] propose a hybrid model, in this case to predict $PM_{2.5}$ and PM_{10} concentrations. Forecasts are made using *multiobjective Harris hawks optimization* and EML. The two objectives to optimize are prediction accuracy and stability of forecasting errors.

E. Conclusion of the Related Works

The study of the state of the art reveals that, although currently the subject of FS for deep learning is being of interest to researchers, it is necessary to deepen in certain aspects. For example, few studies have been carried out on the application of multivariate methods of filter-type FS for deep learning, both for search strategies and evaluators. Regarding wrapper methods for deep learning, although some studies have proposed approaches to reduce computational cost, this remains a challenge for researchers. Finally, the embedded FS methods for deep learning have been compared fundamentally with other embedded methods such as LASSO or decision trees, but not so much with other powerful multivariate FS methods of filter, wrapper, hybrid or surrogate-assisted type. Therefore, in this article we propose and discuss different FS multivariate hybrid methods for deep learning that combine filter methods with wrapper surrogate-assisted methods, using multiobjective evolutionary computation as a search strategy. The proposed methods are compared with a wide range of FS methods that include univariate, multivariate, filter, wrapper, and embedded methods.

The authors of this article have extensive experience in the fields of FS, multiobjective evolutionary computation, time series forecasting, and air quality prediction. In [13] we proposed an MOEA, called evolutionary non-dominated radial slots based algorithm (ENORA), for FS in regression problems. ENORA was used as a search strategy in a wrapper method based on random forest. It was necessary to reduce the number of trees of the random forest algorithm during the optimization process to reduce the computational time of the method. In [47] and [48], MOEAs were used for FS with time series data applied to the prediction of antibiotic resistance outbreaks and the forecast of energy consumption in smart buildings, respectively. In these works, FS wrapper methods based on deep learning were prohibitive due to the excessive computational time required. In [49] deep learning techniques, particularly LSTM, GRU, and CNN, are used for air quality forecasting with time series data. However, in this work FS was not performed, being proposed for future work. Finally, in [50] an MOEA was proposed for the spatio-temporal forecast of air pollution. The proposed technique builds a model based on ensemble learning with the multiple linear regression models found with the MOEA, which is compared with quasi-recurrent neural networks, among other models. Again, FS was raised for future work. Therefore, this article is proposed as a continuation of the previous research, avoiding the inconvenience of computational costs required by deep learning-based FS techniques, by using surrogate-assisted MOEAs.

III. Materials and Methods

In this section, the FS problem is formalized as a multiobjective Boolean optimization problem. The proposed multiobjective optimization models are mathematically defined and the main components of the surrogate-assisted evolutionary algorithm that solve them are described.

A. FS as a Multiobjective Boolean Optimization Problem

FS can be formulated as a *multiobjective boolean optimization problem* as follows:

$$\text{Min./Max. } f_k(\boldsymbol{x}),\ k = 1, \ldots, l \tag{1}$$

where $f_k(\boldsymbol{x})$ are *objective functions* (wrapper type or filter type), $\boldsymbol{x} = \{x_1, x_2, \ldots, x_w\} \in \mathbb{B}^w$ represents the set of *decision variables*, and $\mathbb{B} = \{\text{true, false}\}$ is the domain for each variable $x_i,\ i = 1, \ldots, w$. In problem (1), $x_i = \text{true}$ represents that attribute i is selected, and $x_i = \text{false}$ represents that attribute i is not selected, for all $i = 1, \ldots, w$. Let $X = \{\boldsymbol{x} \in \mathbb{B}^w\}$ be the *search space* of the problem (1). We want to find a subset of solutions $\Psi \subseteq X$ called *non-dominated set* (or *Pareto optimal set*). A solution $\boldsymbol{x} \in X$ is *non-dominated* if there is no other solution $\boldsymbol{x}' \in X$ that dominates $\boldsymbol{x}$, and a solution $\boldsymbol{x}'$ *dominates* $\boldsymbol{x}$ if and only if there exists k $(1 \le k \le l)$ such that $f_k(\boldsymbol{x}')$ improves $f_k(\boldsymbol{x})$, and for every k $(1 \le k \le l)$, $f_k(\boldsymbol{x})$ does not improve $f_k(\boldsymbol{x}')$. In other words, $\boldsymbol{x}'$ *dominates* $\boldsymbol{x}$ if and only if $\boldsymbol{x}'$ is better than $\boldsymbol{x}$ for at least one objective, and not worse than $\boldsymbol{x}$ for any other objective. For minimization problems, the set Ψ of non-dominated solutions of (1) can be formally defined as follows:

$$\Psi = \big\{\boldsymbol{x} \in X \ \mid \ \nexists\, \boldsymbol{x}' \in X \ \mid \ \mathcal{D}\big(\boldsymbol{x}', \boldsymbol{x}\big)\big\}$$

where $\mathcal{D}(\boldsymbol{x}', \boldsymbol{x})$ is equivalent to

$$\exists k,\ 1 \le k \le l,\quad f_k(\boldsymbol{x}') < f_k(\boldsymbol{x}) \\ \wedge \forall k,\ 1 \le k \le l,\quad f_i(\boldsymbol{x}') \le f_i(\boldsymbol{x}).$$

Once the set of optimal solutions is available, the most satisfactory one can be chosen by applying a preference criterion. Search space $X = \{\boldsymbol{x} \in \mathbb{B}^w\}$ of problem (1) contains 2^w solutions (NP-hard problem). Metaheuristics methods such as evolutionary algorithms are typically used to find approximate solutions for NP-hard class problems, including FS problems [51], and they are particularly appropriate for multiobjective optimization.

B. Proposed Multiobjective Optimization Models for FS

Let $D = \{\boldsymbol{d}_1, \ldots, \boldsymbol{d}_s\}$ be a normalized dataset with s instances. Each instance $\boldsymbol{d}_t = \{d_t^1, \ldots, d_t^w, o_t\},\ t = 1, \ldots, s$, has w input attributes of any type, and one output attribute $o_t \in \mathbb{R}$. In this approach, the initial dataset D is divided into three partitions R, V and T for training, validation and test with 60%, 20%, and 20% of the data, respectively. Then these three partitions are normalized. We consider multiobjective boolean optimization models for FS defined as in (1). Let $\mathcal{F}^{\Phi}_{R,V}(\boldsymbol{x})$ be a performance measure, e.g., RMSE, MAE, *correlation coefficient* (CC), and so on, of a regression model built with a learning algorithm Φ trained with dataset $R \subset D$ and evaluated with validation dataset $V \subset D$ using only the selected attributes $x_i = \text{true},\ i = 1, \ldots, w$. Let $\mathcal{C}(\boldsymbol{x})$ be a function that measures the number of selected attributes of $\boldsymbol{x}$, i.e.,

$$\mathcal{C}(\boldsymbol{x}) = \sum_{i=1}^{w} \mathcal{N}(x_i) \tag{2}$$

where $\mathcal{N}$ is a function that transforms a Boolean value into numeric (true $= 1$ and false $= 0$).

Let $\mathcal{P}_R(\boldsymbol{x})$ be a function that measures the sum of correlation between each selected attribute $x_i = \text{true},\ i = 1, \ldots, w$, and the output attribute in dataset R, defined as follows:

$$\mathcal{P}_R(\boldsymbol{x}) = \sum_{\substack{i=1 \\ x_i=\text{true}}}^{w} \rho_i^R \tag{3}$$

where ρ_i^R is the normalized Pearson's correlation coefficient between the selected attribute i and the output in dataset R.

Let $\mathcal{S}_R(\boldsymbol{x})$ be a function that calculates the sum of *reliefF* scores of the selected attributes of $\boldsymbol{x}$ in dataset R, i.e.,

$$\mathcal{S}_R(\boldsymbol{x}) = \sum_{\substack{i=1 \\ x_i=\text{true}}}^{w} \sigma_i^R \tag{4}$$

where σ_i^R is normalized *reliefF* score of attribute i in dataset R. In view of this, we consider the next four objectives.

Objective O1: Minimize the RMSE of the regression model built with learning algorithm Φ trained with dataset R and evaluated with validation dataset V

$$\text{Minimize } O1 \equiv \mathcal{F}^{\Phi}_{R,V}(\boldsymbol{x}).$$

Objective O2: Minimize the number of selected attributes

$$\text{Minimize } O2 \equiv \mathcal{C}(\boldsymbol{x}).$$

Objective O3: Maximize correlation between the selected attributes and the output variable in dataset R

$$\text{Maximize } O3 \equiv \mathcal{P}_R(\boldsymbol{x}).$$

Objective O4: Maximize sum of *reliefF* scores of the selected attributes in dataset R

$$\text{Maximize } O4 \equiv \mathcal{S}_R(\boldsymbol{x}).$$

The following *multiobjective boolean optimization problems* for modeling wrapper-filter FS methods in regression tasks are instances of problem (1) and have been proposed to optimize the objectives defined previously.

1) *Two-Objective Optimization Model for Wrapper FS Method:* Objectives to optimize are *O1*, and *O2*. This problem will be called *O1O2*

$$\begin{aligned} &\text{Minimize} \quad O1 \equiv \mathcal{F}^{\Phi}_{R,V}(\boldsymbol{x}) \\ &\text{Minimize} \quad O2 \equiv \mathcal{C}(\boldsymbol{x}). \end{aligned} \tag{5}$$

2) *Three-Objective Optimization Model for Filter FS Method Based on Correlation and Relief:* Objectives to optimize are *O2*, *O3* and *O4*. This problem will be called *O2O3O4*

$$\begin{aligned} &\text{Minimize} \quad O2 \equiv \mathcal{C}(\boldsymbol{x}) \\ &\text{Maximize} \quad O3 \equiv \mathcal{P}_R(\boldsymbol{x}) \\ &\text{Maximize} \quad O4 \equiv \mathcal{S}_R(\boldsymbol{x}). \end{aligned} \tag{6}$$

3) *Three-Objective Optimization Model for Wrapper-Filter FS Method Based on Correlation:* Objectives to optimize are *O1*, *O2* and *O3*. This problem will be called *O1O2O3*

$$\begin{aligned} &\text{Minimize} \quad O1 \equiv \mathcal{F}^{\Phi}_{R,V}(\boldsymbol{x}) \\ &\text{Minimize} \quad O2 \equiv \mathcal{C}(\boldsymbol{x}) \\ &\text{Maximize} \quad O3 \equiv \mathcal{P}_R(\boldsymbol{x}). \end{aligned} \tag{7}$$

4) *Three-Objective Optimization Model for Wrapper-Filter FS Method Based on ReliefF:* Objectives to optimize are *O1*, *O2* and *O4*. This problem will be called *O1O2O4*

$$\begin{aligned} &\text{Minimize} \quad O1 \equiv \mathcal{F}^{\Phi}_{R,V}(\boldsymbol{x}) \\ &\text{Minimize} \quad O2 \equiv \mathcal{C}(\boldsymbol{x}) \\ &\text{Maximize} \quad O4 \equiv \mathcal{S}_R(\boldsymbol{x}). \end{aligned} \tag{8}$$

5) *Four-Objective Optimization Model for Wrapper-Filter FS Method Based on Correlation and ReliefF:* Objectives to optimize are *O1*, *O2*, *O3*, and *O4*. This problem will be called *O1O2O3O4*.

$$\begin{aligned} &\text{Minimize} \quad O1 \equiv \mathcal{F}^{\Phi}_{R,V}(\boldsymbol{x}) \\ &\text{Minimize} \quad O2 \equiv \mathcal{C}(\boldsymbol{x}) \\ &\text{Maximize} \quad O3 \equiv \mathcal{P}_R(\boldsymbol{x}) \\ &\text{Maximize} \quad O4 \equiv \mathcal{S}_R(\boldsymbol{x}). \end{aligned} \tag{9}$$

C. Surrogate-Assisted MOEA

The function $\mathcal{F}^{\Phi}_{R,V}(\boldsymbol{x})$ measures the performance (we use RMSE) of a prediction model training with learning algorithm Φ (we use LSTM) on dataset R with attributes of $\boldsymbol{x}$ and evaluated on validation dataset V. $\mathcal{F}^{\Phi}_{R,V}(\boldsymbol{x})$ can be used as an evaluator in a wrapper FS method. As is well known, the drawback of wrapper methods for FS is the computational time required by the learning algorithm to build the models with the attribute subsets explored with the search strategy. This disadvantage is increased in the presence of large datasets or when the learning algorithm is particularly expensive, as is the case with deep learning. To take advantage of the good properties of wrapper methods without the need to train the model on each subset of candidate attributes, in this article we propose a *surrogate-assisted* approach. The set of attributes selected by the surrogate-assisted MOEA proposed in this article is finally evaluated on the test set T, which has not been seen by the MOEA. Fig. 1 shows the flowchart of the proposed surrogate-assisted MOEA, and its main components are shown below.

1) Representation of Solutions: Our MOEA uses a fixed-length representation, where each individual consists of a bit vector for attribute selection. Therefore, an individual I is represented as follows:

$$I = \{b_1^I, \ldots, b_w^I\}$$

where $b_i^I \in \{0, 1\}$ for $i = 1, \ldots, w$. Each bit $b_i^I \in \{0, 1\}$ represents an attribute in the dataset (1 for selected, and 0 for non-selected attributes).

2) Initial Population: The initial population is randomly generated with a Bernoulli distribution with probability 0.5.

3) Fitness Function: We use the surrogate function $\mathcal{Q}^{\Phi}_{R,V}(\boldsymbol{x})$ as an approximation to the function $\mathcal{F}^{\Phi}_{R,V}(\boldsymbol{x})$ for the calculation of the objective function *O1*. $\mathcal{Q}^{\Phi}_{R,V}(\boldsymbol{x})$ is the RMSE of a surrogate prediction model M^{Φ}_R trained with learning algorithm Φ (we use LSTM) and dataset R considering all w attributes, and evaluated with a dataset V' which consists of the validation dataset V where attribute values x_i = false, $i = 1, \ldots, w$, are set to a constant value α in all s instances (we used

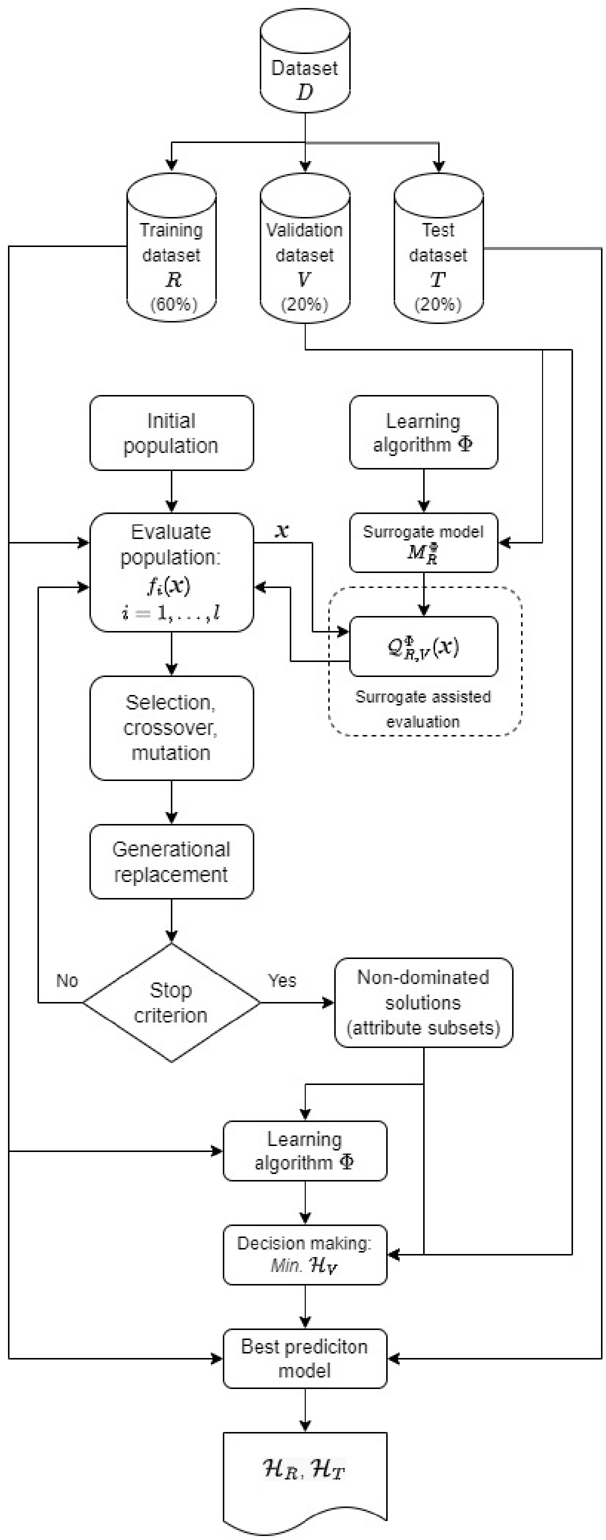

Fig. 1. Flowchart of the proposed surrogate-assisted MOEA.

$\alpha = 0$ in the experiments, although other values can also be used). Note that the surrogate model M_R^Φ does not depend on the set $\boldsymbol{x}$ of decision variables, so it is trained only once in off-line mode. Note also that $\mathcal{Q}_{R,V}^\Phi(\boldsymbol{x}_D) = \mathcal{F}_{R,V}^\Phi(\boldsymbol{x}_D)$, where $\boldsymbol{x}_D$ is the solution vector with $x_i =$ true, for all $i = 1, \ldots, w$. Algorithm 1 shows the calculation of the fitness function for

Algorithm 1 Fitness Function for Objective Function *O1*

Require: $I = \{b_1^I, \ldots, b_w^I\}$ {Individual}
Require: M_R^Φ {Surrogate prediction model built with learning algorithm Φ and trained with dataset R with all attributes}
Require: $V \subset D$ {Validation dataset}
Require: α {Imputation constant}
$V' \leftarrow V$
for $i = 1$ **to** w **do**
  if $b_i^I = 0$ **then**
    for $\boldsymbol{d}_t \in V'$ **do**
      $d_t^i = \alpha$
    end for
  end if
end for
return $RMSE(M_R^\Phi, V')$ {RMSE of surrogate prediction model M_R^Φ evaluated in dataset V'}

objective *O1*. The fitness functions for the rest of the objective functions *O2*, *O3*, and *O4* correspond to the mathematical formulations described in (2)–(4), respectively.

4) Variation Operators: The variation operators used are *half uniform crossover* [52] and *bit flip mutation* [53]. We have chosen these operators since they are the ones used by the Platypus platform (https://platypus.readthedocs.io/en/latest/) for MOEAs with binary representation.

5) Decision Process to Choose the Final Non-Dominated Solution: For each non-dominated individual in the last population, an LSTM model m is trained on dataset R with the attributes selected in the individual, and a *recursive multistep strategy* [54] is performed for h-steps ahead. Although there are other strategies for multistep ahead time series forecasting [55], [56], in this research we have decided to use the recursive strategy. The recursive strategy constrains all the horizons to be forecast with the same model structure, and can suffer from poor performance in multistep forecasting tasks, especially if the forecast horizon h exceeds the window size l, since at some point all the inputs are forecast values rather than actual observations. Nevertheless, this does not occur in our research since we use $h \leq l$. In spite of these drawbacks, we have preferred to adopt the recursive strategy as a first approximation of our work since it is not computationally expensive and can be implemented in all the forecasting methods compared in this article.

The non-dominated individual is evaluated on dataset V using a performance measure $\mathcal{H}$. The chosen non-dominated individual will be the one with the lowest value of $\mathcal{H}$. Therefore, an ideal solution is one with $\mathcal{H} = 0$. Algorithm 2 calculates the performance measure $\mathcal{H}$ of a solution $\boldsymbol{x}$ on a test dataset E, which is particularly used to evaluate the non-dominated individuals for MOEA decision making by setting $E = V$.

IV. Experiments and Results

This section describes the air quality dataset used in the experiments and its preprocessing. The experiments have been established for the comparison of the different multiobjective

Algorithm 2 Evaluation $\mathcal{H}$ of a Solution on a Dataset E

Require: $\boldsymbol{x} = \{x_1, \ldots, x_w\}$ {Solution}
Require: m {Prediction model trained with the attributes selected in the solution $\boldsymbol{x}$}
Require: E {Dataset for evaluation}
Require: h {Number of steps-ahead}
Require: $w_1, w_2, w_3,\ 0 \leq w_1, w_2, w_3 \leq 1,\ w_1 + w_2 + w_3 = 1$ {Weights}

$E' \leftarrow E$
for $i = 1$ **to** w **do**
  if $x_i = 0$ **then**
    $remove(E', i)$ {Remove attribute i from dataset E'}
  end if
end for
$RMSE \leftarrow \frac{1}{h}\sum_{j=1}^{h} RMSE(m, j, E')$
$MAE \leftarrow \frac{1}{h}\sum_{j=1}^{h} MAE(m, j, E')$
$CC \leftarrow \frac{1}{h}\sum_{j=1}^{h} \left|1 - CC(m, j, E')\right|$
$\mathcal{H} \leftarrow w_1 \cdot RMSE + w_2 \cdot MAE + w_3 \cdot CC$
return $\mathcal{H}$ **{Evaluation of a solution}**

Fig. 2. Photograph of the air quality measurement station in La Aljorra. Google image.

optimization models for FS proposed in this article, the comparison of the MOEAs that solve them, the identification and adjustment of the best prediction model and the comparison with other FS techniques.

A. Air Quality Dataset

Data are extracted from the *autonomous community of the region of Murcia*, Spain (https://sinqlair.carm.es/calidadaire/redvigilancia/redvigilancia.aspx), which provides information on air quality in the Region of Murcia thanks to *Consejería de Agua, Agricultura, Ganadería, Pesca y Medio Ambiente*.

All the collected data comes from La Aljorra station (Fig. 2), taken daily for four years, from 2017 to 2020. In total, there are 1461 instances. The initial dataset consisted of 17 columns: *Date*, *NO*, NO_2, SO_2, O_3, *TMP (temperature)*, *HR (relative humidity)*, NO_X, *DD (wind direction)*, *PRB (atmospheric pressure)*, *RS (solar radiation)*, *VV (wind speed)*, C_6H_6, C_7H_8, *XIL*, PM_{10}, *Noise*. Table I shows a summary of the values of all attributes.

TABLE I
SUMMARY OF INITIAL ATTRIBUTES

Name	Units	Count	Mean	Std	Min	Max
Date	-	1461	-	-	-	-
NO	$\mu g/m^3N$	1272	4.38	2.43	1.00	31.00
SO_2	$\mu g/m^3N$	1299	9.21	3.45	2.00	23.00
O_3	$\mu g/m^3N$	1403	57.94	15.25	19.00	112.00
TMP	*ºC*	1409	19.59	5.74	4.00	32.00
HR	*% R.H.*	1409	69.21	14.09	27.00	100.00
NO_X	$\mu g/m^3N$	1272	21.21	11.20	3.00	104.00
DD	*degrees*	1409	192.11	104.45	0.00	360.00
PRB	*mb*	1409	1017.81	6.41	991.00	1033.00
RS	W/m^3	1409	182.32	82.74	13.00	338.00
VV	m/s	1409	1.22	0.43	1.00	3.00
C_6H_6	$\mu g/m^3N$	410	0.65	0.51	0.10	3.40
C_7H_8	$\mu g/m^3N$	410	0.93	0.40	0.30	3.10
XIL	$\mu g/m^3N$	410	1.40	0.68	0.10	3.00
PM_{10}	$\mu g/m^3N$	1381	26.27	12.98	5.00	168.00
Noise	*dBA*	0	-	-	-	-
NO_2 (target)	$\mu g/m^3N$	1272	14.64	8.19	2.00	58.00

B. Data Preprocessing: Missing Values Imputation and Sliding Window Transformation

The proposed dataset is used for NO_2 prediction. For the initial preprocessing of the data, those columns with more than 25% missing values were removed. These columns were: C_6H_6, C_7H_8, *XIL* and *Noise*. *Date* column has also been eliminated, as it does not provide any relevant information to the treated problem. The remaining missing values have been imputed with linear interpolation.

Lagged transformation of the input variables has to be done with *sliding window* method [57] in order to remove time dependencies in the data. Let l be the *window size* (WS), i.e., the number of previous time steps. The sliding window transformation process builds the following dataset D_l:

$$D_l = \left\{\left\{d^1_{t-1}, \ldots, d^1_{t-l}\right\}, \ldots, \left\{d^w_{t-1}, \ldots, d^w_{t-l}\right\} \\ \{o_{t-1}, \ldots, o_{t-l}\}, o_t\right\}, \quad t = 1, \ldots, s. \qquad (10)$$

Note that $d^i_t, i = 1, \ldots, w$, and o_t values with $t \leq 0$ do not exist and are therefore considered missing values. The new attributes are named *<Lag>_<name of the original attribute>_<number of previous time step>*. For example, *NO* attribute with 1 previous time step will be called *Lag_NO_1*. A window size of 7 has been selected, representing one week. The rows with missing values resulting from the sliding window transformation have been removed. Note that the original variables d^i_t, $i = 1, \ldots, w$, of the input attributes have been removed, as it is not correct to use the future value of an attribute to obtain future predictions. The only original variable held is the output. Therefore, our dataset with WS 7 has 1454 rows and 85 columns. Finally, the dataset has been split into 60% for training, 20% for validation and 20% for testing, preserving the order of the instances and then normalized.

C. Multiobjective Optimization Models Comparison

This section describes the process of searching for the best hyper-parameters of the surrogate deep learning model, the evaluation of the prediction models for each multiobjective optimization model and the statistical tests applied to select the best model.

1) Hyper-Parameter Tuning of the Surrogate-Assisted Model: For the creation of the M_R^Φ surrogate prediction model used in the evolutionary multiobjective algorithm we have selected LSTM. Our previous study [49] showed that an LSTM with one hidden layer with *rectified linear unit* [58] activation function and a dense layer with a linear activation as the output layer can obtain good results in time series forecasting applied to air quality prediction, over other learning algorithms such as 1-D-CNN, random forest, SVM or LASSO regression. In order to optimize this model, we first search for the best hyper-parameters. For this purpose, a threefold cross-validation with one repetition was used. A grid search has been carried out with: one, two, and three hidden layers; 50, 100, 150, 200, 250, and 300 neurons; a batch size of 32, 64, 128, 512, and 1024; and 100, 500, and 1000 epochs. Therefore, 270 possible combinations have been performed. The best combination of hyper-parameters obtained an RMSE of 0.07877 with a standard deviation of 0.01339 and was one hidden layer, 200 neurons, batch size of 128 and 100 epochs. The LSTM architecture used as a surrogate model in this article consists of an input layer, a hidden layer with 200 neurons, and ReLU activation function, a layer with a dropout of 0.2 and an output layer with one neuron and linear activation function.

A surrogate LSTM model M_R^Φ has been trained with all attributes, a 60% of the instances and with the best hyper-parameters, and tested with 20% of the validation instances. The RMSE in the training dataset is 0.0515 and the RMSE in the validation dataset is 0.1795.

2) Prediction Model Evaluation for Each Multiobjective Optimization Model: For our dataset, the proposed multiobjective optimization problems (5)–(9) have been solved with NSGA-II. NSGA-II is widely recognized by the scientific community as a representative algorithm for solving multiobjective optimization problems. This MOEA is used in this initial phase of the experiments to compare the different multiobjective optimization models, although in a later experimentation phase (Section IV-D) NSGA-II is compared with a broader set of MOEAs. It has been run 10 times for each problem, with a population size of 100, 1000 generations (100 000 evaluations of the objective functions) and crossover and mutation probabilities of 1.0. Experiments were run on a computer with an AMD Ryzen 7 5800X 8-Core Processor with 3.80 GHz using 16 GB of RAM and Windows 10 Pro. We use RMSE as performance metric $\mathcal{Q}_{R,V}^\Phi(\boldsymbol{x})$. For the decision-making process in order to choose the best non-dominated solution (Algorithm 2 with the validation dataset V) we have used multistep ahead predictions and normalized weights $w_1 = w_2 = w_3$. Finally, the best model obtained with NSGA-II is evaluated by using Algorithm 2 with the (unseen) test dataset T.

Table II shows for each optimization problem, the results of the best solution including the evaluation $\mathcal{H}$ on training, validation and test datasets ($\mathcal{H}_R$, $\mathcal{H}_V$ and $\mathcal{H}_T$, respectively) and the number of selected attributes. Therefore, we will have six datasets: five reduced datasets selected by the NSGA-II algorithm and the original dataset.

TABLE II

PREDICTION MODEL EVALUATION WITH NSGA-II FOR THE AIR QUALITY PROBLEM, 100 000 EVALUATIONS AND TEN RUNS, SORTED FROM BEST TO WORSE EVALUATION OF $\mathcal{H}_T$

Optimization model	$\mathcal{Q}_{R,V}^\Phi(\boldsymbol{x})$	Number of selected attributes	$\mathcal{H}_R$	$\mathcal{H}_V$	$\mathcal{H}_T$	Run time (minutes)
O1O2O3	0.1939	16	0.0807	0.2182	0.1298	15.76
O1O2	0.2403	2	0.1234	0.2328	0.1442	9.80
O1O2O3O4	0.2006	4	0.1210	0.2347	0.1447	28.41
O1O2O4	0.2462	6	0.0928	0.2447	0.1647	18.33
O2O3O4	–	19	0.1006	0.2746	0.1965	7.65

TABLE III

RANKING OF MULTIOBJECTIVE OPTIMIZATION MODELS, TENFOLDS CROSS-VALIDATION, THREE REPETITIONS, SORTED FROM BEST TO WORSE

Optimization model	Win	Loss	Win-Loss
O1O2O3	4	0	4
O1O2O4	3	1	2
O1O2	1	2	-1
O2O3O4	1	2	-1
O1O2O3O4	0	4	-4

3) Statistical Tests for Performance Prediction Models: Once the best reduced dataset for each optimization problem has been found, three repetitions of tenfold cross-validation with the training dataset R (a total of 30 prediction models) are performed using the Algorithm 2 for evaluation. In order to apply statistical tests, first, we check whether the conditions of normality and sphericity of the samples are met. For normality, Shapiro-Wilks test was used, if p-values are greater than 0.05 then it can be assumed that the values come from a normal distribution. For sphericity, we used Mauchly's test, if p-values are greater than 0.05 then the sphericity condition is satisfied. Then, tests are applied to check if there are statistically significant differences between the models. Parametric ANOVA test in case the normality and sphericity conditions are met and non-parametric Friedman test in case they are not met. If the Friedman test is applied, Nemenyi post hoc test is also used to determine where the significant differences between the models are (those with a p-value less than 0.05).

In order to summarize results and determine the best and worst models, method-dataset pairs have been ranked according to the difference between *wins* and *losses* obtained by Wilcoxon signed-rank test. Every time one method-dataset pair tests statistically significantly better than another, it counts as a *win* and otherwise as a *loss*. Table III shows the ranking of every multiobjective optimization model.

D. MOEAs Comparison

1) MOEAs Performance Evaluation: Once the best optimization model has been identified, the best MOEA is determined in this new phase of experiments. We have selected the MOEAs implemented in the Platypus platform that allow binary representations, which are NSGA-II, NSGA-III, *MOEA based on decomposition* (MOEA/D) [59], *indicator-based evolutionary algorithm* (IBEA) [60], ϵ-MOEA [61], *strength Pareto evolutionary algorithm 2* (SPEA2) [62] and

TABLE IV

RANKING OF THE OPTIMIZATION ALGORITHMS WITH HYPERVOLUME METRIC, 100 000 EVALUATIONS, 30 RUNS, SORTED FROM BEST TO WORSE

MOEA	Win	Loss	Win-Loss
NSGA-II	6	0	6
IBEA	4	1	3
ϵ-MOEA	4	1	3
ϵ-NSGA-II	3	3	0
NSGA-III	2	4	-2
SPEA2	1	5	-4
MOEA/D	0	6	-6

ϵ-NSGA-II [63]. Each algorithm is run 30 times with the best multiobjective model (*O1O2O3*) with 100 000 evaluations and probability 1.0 of crossover and mutation. Afterward, *hypervolume* metric is calculated. A set of *reference points* is required for the calculation of a hypervolume metric. For the multiobjective optimization model *O1O2O3*, the set of reference points consists of 84 points (as many as attributes in the original dataset) in a 3-D space (1-D for each objective). Dimension *O1* with 0 attributes is the value of $\mathcal{Q}^{\Phi}_{R,V}(\boldsymbol{x}_D)$ model when no attribute is selected, i.e., 0.2647 in our case. All other values of that dimension are 0. Dimension *O2* indicates the number of attributes, so it will range from 0 to 84. Dimension *O3* is the sum of the list ordered from best to worst correlation. That is, for 0 attributes it will be 0, for 1 attribute it will be the best correlation, for 2 attributes it will be the sum of the two best correlations and so on.

2) Statistical Tests for MOEA's Performance: The statistical test Mann–Whitney U rank is applied in order to rank the algorithms and select the best. Table IV shows the win-loss ranking for MOEAs.

E. Final Prediction Model Identification and Adjustment

For each metric, those MOEAs that have never lost in any of the statistical tests are selected. The chosen algorithm is NSGA-II. Therefore, the best model is found with *O1O2O3* optimization model with NSGA-II algorithm with crossover and mutation probabilities 1.0.

1) Hyper-Parameter Tuning of the MOEA: A hyper-parameter tuning is carried out on NSGA-II to establish the best crossover and mutation probabilities. The probabilities of both crossover and mutation on which the search has been performed have been 0.2, 0.5, 0.8, and 1.0 with ten runs and 10 000 evaluations. The best combination was a crossover probability of 0.2 and a mutation probability of 0.2 for the hypervolume metric.

2) Model Adjustment: The best MOEA (NSGA-II) is rerun for the best optimization model (*O1O2O3*) with the seed that produced the best result in the previous phase of experiments, and the best probabilities obtained for the hypervolume metric (crossover probability 0.2 and mutation probability 0.2) with 1 000 000 evaluations. From the last population, the non-dominated solutions are obtained, the models are trained with LSTM and the average of the seven-steps ahead predictions are calculated. The evaluation $\mathcal{H}$ on the training dataset R, the validation dataset V and the test dataset T are $\mathcal{H}_R = 0.1234$, $\mathcal{H}_V = 0.2328$, and $\mathcal{H}_T = 0.1441$, respectively. Note that the prediction model obtained using crossover probability 0.2 and mutation probability 0.2 does not improve the prediction model obtained using crossover and mutation probability 1.0. This is because the hypervolume metric measures the optimality and diversity of the MOEA based on the ideal population, but this metric does not explicitly measure the performance of the individuals as defined in Algorithm 2. Therefore, the best model with probability 1.0 of crossover and mutation is considered.

F. Comparison With Other FS Techniques

In this section, we compare our proposed solution with other FS techniques of different types. Among them, there are two filter-wrapper hybrid FS methods that combine attribute evaluation methods with LSTM and deterministic search, based on correlation and the reliefF algorithm, respectively. We also compare our method with two wrapper multiobjective evolutionary FS methods (without surrogates) based on linear regression and random forest, respectively, which solve the *O1O2* multiobjective optimization model using NSGA-II. Finally, we compare two embedded FS methods: CancelOut, implemented in an LSTM, and random forest.

M1) Hybrid Filter-Wrapper FS Method Based on Correlation and LSTM With Deterministic Search: This method , also proposed by the authors of this article, uses the normalized Pearson's correlation coefficient between the selected attribute i and the output to sort all attributes from highest to lowest correlation. Then, w LSTM models are built with the i best attributes in the ranking, for all $i = 1, \ldots, w$. The attribute subset that produces the LSTM model with the lowest RMSE is selected (Algorithm 3). This method is a hybrid between filter and wrapper since it uses together a correlation filter and an LSTM-based learning algorithm. It is also a multivariate FS method since although the attributes are ordered by their individual correlation with the output, attribute subsets are then formed with the most correlated attributes, in a deterministic way with a total of w candidate subsets, which are evaluated by means of an LSTM detecting, therefore, interactions between factors. Note that this method is equivalent to solving the multiobjective optimization problem formed by objectives *O2* and *O3*, whose Pareto front consists of the set of solutions $\Psi = \{s_i \in \mathbb{B}^w, i = 1, \ldots, w\}$ obtained with Algorithm 3, to finally select the solution s_i that builds the best prediction model m_{best}. However, the *O2O3* multiobjective optimization problem can be solved deterministically in this case with Algorithm 3 without resorting to approximate probabilistic methods such as MOEAs.

M2) Hybrid Filter-Wrapper FS Method Based on reliefF and LSTM With Deterministic Search: This method is similar to the previous one except that it uses the multivariate reliefF filter instead of the univariate correlation filter, and has also been proposed by the authors of this article. Algorithm 3 therefore also serves to describe this FS method. This method also uses a hybrid wrapper-filter technique, but in this case, it can be considered a fully multivariate method. This method is equivalent to solving the multiobjective optimization problem with the objectives *O2* and *O4* and selecting the best

Algorithm 3 Hybrid Filter-Wrapper FS Method Based on Correlation/reliefF and LSTM With Deterministic Search

Require: $\boldsymbol{r} = \{r_1, \ldots, r_w\}$, $1 \le r_i \le w \mid \rho^R_{r_{i-1}} \ge \rho^R_i, \forall i = 2, \ldots, w$ $\{\sigma^R_{r_{i-1}} \ge \sigma^R_i,\ \forall i = 2, \ldots, w$, for reliefF$\}$
Require: $R \subset D$ {Training dataset}
Require: $V \subset D$ {Validation dataset}
Require: h {Number of steps-ahead}
Require: w_1, w_2, w_3, $0 \le w_1, w_2, w_3 \le 1$, $w_1 + w_2 + w_3 = 1$ {Weights}
for $i = 1$ **to** w **do**
  $x_i = 0$
end for
for $i = 1$ to w **do**
  $x_{r_i} \leftarrow 1$
  $m_i \leftarrow \text{LSTM}(R, \boldsymbol{x})$ {Prediction model built with LSTM in dataset R with the attribute subset $\boldsymbol{x}$}
  $\boldsymbol{s}_i \leftarrow \boldsymbol{x}$
end for
$best \leftarrow \arg\min_{i=1,\ldots,w} \mathcal{H}(\boldsymbol{s}_i, m_i, V, h, w_1, w_2, w_3)$ {Using Algorithm 2}
return m_{best}

solution from the Pareto front, which can also be solved deterministically with Algorithm 3.

M3) Wrapper Multiobjective Evolutionary FS Method Based on Linear Regression: This is a multivariate wrapper method, for evaluating subsets of attributes, that uses NSGA-II with search strategy and a linear regression algorithm as evaluator together with the RMSE metric. The multiobjective optimization model *O1O2* described in equation (5) is solved. This method does not use surrogates but rather a linear regression model is trained for each candidate attribute subset. It has been evaluated with ten runs, 100 000 evaluations and 1.0 crossover and mutation probabilities. Each set of selected attributes is evaluated with a linear regression model.

M4) Wrapper Multiobjective Evolutionary FS Method Based on Random Forest: This FS method is similar to the previous one except that the random forest learning algorithm is used for the evaluation of attribute subsets.

M5) CancelOut: CancelOut [23] is an embedded method that adds a new input layer for ANNs. Each neuron in this layer is connected to an input variable. The weights are updated during a training stage so that irrelevant features are "cancelled" and those that do provide information are maintained. The CancelOut layer has three parameters: an activation function, λ_1 and λ_2. λ_1 and λ_2 are two regularization terms to speed up the FR process. The values of both lambdas are between 0 and 1. A search has been performed to find the best CancelOut hyper-parameters. For this purpose, the CancelOut layer has been added to the previously described LSTM architecture. The values of λ_1 and λ_2 on which the search will be done are: 0, 0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5, and 1. The best value for λ_1 was 0.5 and for λ_2 was 0.05. The activation function will always be the sigmoid, which is the default parameter proposed by the authors of the layer. The LSTM architecture consists of an input layer, followed by a CancelOut layer, a hidden layer with 200 neurons and ReLU activation function, a layer with a dropout of 0.2 and an output layer with one neuron and linear activation function.

TABLE V
COMPARISON OF FS METHODS FOR THE AIRQUALITY PROBLEM, SORTED FROM BEST TO WORSE EVALUATION OF $\mathcal{H}_T$

Method	$\mathcal{H}_R$	$\mathcal{H}_T$	Number of selected attributes	Run time (minutes)
O1O2O3-NSGA-II	0.0807	0.1298	16	15.76
M1	0.1246	0.1437	2	3.93
M2	0.1246	0.1437	2	4.09
M4	0.1235	0.1876	6	22.60
M6	0.1701	0.2069	1	2.16
M3	0.1227	0.2243	14	5.45
M5	0.0602	0.2452	84	0.07
All attributes	0.0560	0.2763	84	0.01

TABLE VI
RESULTS OF THE BEST PREDICTION MODEL (OBTAINED WITH *O1O2O3-NSGA-II* METHOD) FOR THE AIR QUALITY PROBLEM EVALUATED ON R AND T DATASETS

Set	Metric	1-step ahead	2-steps ahead	3-steps ahead	4-steps ahead	5-steps ahead	6-steps ahead	7-steps ahead
R	RMSE	0.0761	0.0744	0.0749	0.0751	0.0755	0.0758	0.0773
	MAE	0.0533	0.0529	0.0534	0.0535	0.0537	0.0538	0.0548
	CC	0.8892	0.8916	0.8900	0.8885	0.8861	0.8846	0.8805
T	RMSE	0.0814	0.0819	0.0822	0.0829	0.0832	0.0848	0.0873
	MAE	0.0494	0.0498	0.0503	0.0505	0.0507	0.0517	0.0537
	CC	0.7535	0.7508	0.7507	0.7478	0.7467	0.7380	0.7257

M6) Random Forest: Random forest is also an embedded method that performs FS internally. To do so, random forest considers how each attribute affects the impurity of the nodes.

In Table V, the compared FS techniques are sorted from best to worst evaluation of $\mathcal{H}$ on the test dataset T ($\mathcal{H}_T$). The run times are in minutes. In addition, the original LSTM model trained with all attributes has also been included in the comparison. Table VI shows the RMSE, MAE, and CC of the seven-steps ahead predictions of the best prediction model (*O1O2O3-NSGA-II*) evaluated on the training dataset R and the test dataset T.

G. Additional Experiments: Indoor Temperature Forecasting in a Domotic House

In this section, we analyze a second forecasting problem in order to strengthen the conclusions about the proposal. The forecast of the indoor temperature in a domotic house in Valencia (Spain) is now considered. The dataset (*SML2010*) has been extracted from the *UCI machine learning repository* [64]. In its original version [65], the dataset contains times series with a total of 24 attributes and 4137 instances, with data acquired in Valencia (Spain) from 03/13/2012 to 05/02/2012, each 15 min. In this study, we have set the indoor temperature of the dining room as the target. After removing date and time attributes, useless attributes, and attributes related to the indoor temperature of another room, the set of 15 attributes shown in Tables VII and VIII is obtained.

Applying the methodology followed in this article with a WS of 4, we obtain that the best optimization model has been *O1O2O3O4* in this case (Table IX). Note that the sliding window has not been applied to the DW variable, due to its categorical character. The best MOEA has turned out to

TABLE VII
ATTRIBUTES IN SML2010 DATASET USED IN THIS ARTICLE

Short name	Description
WT	Weather forecast temperature, in ^oC
CO_2	Carbon dioxide, in *ppm* (dinning room)
RH	Relative humidity (dinning room), in $\%$
L	Lighting (dinning room), in *Lux*
R	Rain, in range $[0, 1]$
SD	Sun dusk
W	Wind, in *m/s*
SLW	Sun light in west facade, in *Lux*
SLE	Sun light in east facade, in *Lux*
SLS	Sun light in south facade, in *Lux*
SI	Sun irradiance, in W/m^2
OT	Outdoor temperature, in ^oC
ORH	Outdoor relative humidity, in $\%$
DW	Day of the week, 1 = Monday, 7 = Sunday
IT (target)	Indoor temperature (dinning-room), in ^oC (target)

TABLE VIII
SML2010 ATTRIBUTE STATISTICS

Name	Mean	Std	Min	Max
WT	15.09	4.38	0.00	29.00
CO_2	206.60	22.76	187.34	594.39
RH	42.39	7.22	26.17	60.96
L	28.97	25.68	10.74	111.80
R	0.04	0.19	0.00	1.00
SD	335.09	304.51	0.61	625.00
W	1.30	1.22	0.00	6.32
SLW	14749.15	25306.45	0.00	95278.40
SLE	13566.28	23311.85	0.00	92367.50
SLS	19857.18	29494.60	0.00	95704.40
SI	232.20	312.46	-4.16	1094.66
OT	18.02	4.29	9.22	29.91
ORH	53.25	13.51	22.25	83.81
DW	3.96	1.99	1.00	7.00
IT (target)	20.49	3.31	11.35	28.92

be NSGA-II also in this problem. The *O1O2O3O4-NSGA-II* prediction model has been the best compared to the rest of the models obtained with the FS methods described in Section IV-F. Table X shows the results of the comparison. Table XI shows the RMSE, MAE, and CC of the four-steps ahead predictions of the *O1O2O3O4-NSGA-II* model evaluated on the training dataset R and the test dataset T.

TABLE IX
PREDICTION MODEL EVALUATION WITH NSGA-II FOR THE INDOOR TEMPERATURE PROBLEM, 100 000 EVALUATIONS, TEN RUNS, SORTED FROM BEST TO WORSE EVALUATION OF $\mathcal{H}_T$

Optimization model	$\mathcal{Q}^{\Phi}_{R,V}(\boldsymbol{x})$	Number of selected attributes	$\mathcal{H}_R$	$\mathcal{H}_V$	$\mathcal{H}_T$	Run time (minutes)
O1O2O3O4	0.0364	10	0.0061	0.0091	0.0115	38.03
O1O2O3	0.0449	21	0.0075	0.0105	0.0130	27.18
O1O2	0.0659	4	0.0114	0.0126	0.0136	13.24
O1O2O4	0.0659	4	0.0114	0.0126	0.0136	32.34
O2O3O4	–	26	0.0062	0.0101	0.0149	14.43

TABLE X
COMPARISON OF FS METHODS FOR THE INDOOR TEMPERATURE PROBLEM, SORTED FROM BEST TO WORSE EVALUATION OF $\mathcal{H}_T$

Method	$\mathcal{H}_R$	$\mathcal{H}_T$	Number of selected attributes	Run time (minutes)
O1O2O3O4-NSGA-II	0.0061	0.0115	10	38.03
M6	0.0079	0.0130	6	2.87
M1	0.0075	0.0141	28	5.08
All attributes	0.0112	0.0168	57	0.10
M5	0.0112	0.0168	57	0.09
M2	0.0089	0.0219	9	5.13
M4	0.0321	0.0394	1	29.44
M3	0.0321	0.0394	1	9.12

TABLE XI
RESULTS OF THE BEST PREDICTION MODEL (OBTAINED WITH *O1O2O3O4-NSGA-II* METHOD) FOR THE INDOOR TEMPERATURE PROBLEM EVALUATED ON R AND T DATASETS

Set	Metric	1-step ahead	2-steps ahead	3-steps ahead	4-steps ahead
R	RMSE	0.0079	0.0094	0.0108	0.0126
	MAE	0.0052	0.0064	0.0075	0.0089
	CC	0.9993	0.999	0.9987	0.9983
T	RMSE	0.0129	0.016	0.0194	0.0235
	MAE	0.0112	0.0139	0.0169	0.0204
	CC	0.9994	0.9991	0.9988	0.9983

V. ANALYSIS OF RESULTS AND DISCUSSION

This section provides an analysis of the results presented above attending to the comparison of the multiobjective optimization models and the MOEAs that solve them, as well as the identification and adjustment of the final prediction model and its comparison with prediction models obtained with other FS techniques.

A. Analysis of the Multiobjective Optimization Models

The experiments aimed at identifying the best multiobjective optimization model have been based on the evaluation of the prediction models obtained and on statistical tests to discard those models that have presented statistically significant differences. We have used the performance measure $\mathcal{H}$, proposed in this article, to evaluate the prediction models on a dataset E (Algorithm 2, where E has been the dataset R, V or T) taking into account the RMSE, MAE, and CC indicators in multistep ahead predictions. For the statistical tests, the prediction models have been evaluated by means of three repetitions of ten-fold cross-validation, therefore using a total of 30 prediction models for each multiobjective optimization model. The following analysis can be derived as follows.

1) The combination of the evaluation of the surrogate model together with the correlation filter (multiobjective optimization model *O1O2O3*) produces the best prediction model (evaluated on the T dataset) in the air quality problem, while in the indoor temperature problem, the best resulting optimization model combines the evaluation of the surrogate model with the correlation and reliefF filters (multiobjective optimization model *O1O2O3O4*). This indicates that all the objective functions defined in this article can be important in the multiobjective search for attribute subsets in the problems under study. Statistical tests have confirmed the superiority of these optimization models in their respective application problems.
2) The *O1O2O3* optimization model for air quality forecasting has selected 16 attributes, and thus has removed 68 attributes from the original dataset (out of 84 attributes), as well as improving the performance of the LSTM neural network. The *O1O2O3O4*

optimization model for indoor temperature forecasting has selected 10 attributes, removing 47 attributes from the original dataset (out of 57 attributes) and also improved the performance of the LSTM neural network.

3) With the *O1O2* optimization model, which does not contain the correlation filter or the reliefF filter, a lower number of attributes is selected than in the rest of the optimization models, in both air quality and indoor temperature problems. However, the performance of the prediction model obtained deteriorates with respect to other models that do contain some of these filters.
4) The multiobjective optimization models *O1O2* and *O2O3O4* have the shortest run times, while the multiobjective optimization model *O1O2O3O4* containing all objectives has, as expected, the longest run time.

B. MOEAs Comparison

In order to identify which MOEA has better performance, several outstanding MOEAs from the literature have been compared using optimality and diversity metrics used in multiobjective optimization. The multiobjective optimization models *O1O2O3*, for the air quality problem, and *O1O2O3O4*, for the indoor temperature problem, have been used in the comparison of MOEAs. Again, statistical tests have been performed to select, in this case, those MOEAs that have never lost in any of the statistical tests. For both multiobjective optimization models *O1O2O3* and *O1O2O3O4* in their respective application problems, the NSGA-II algorithm turned out to be the best. Figs. 3 and 4 show the Pareto fronts found with *O1O2O3-NSGA-II* for air quality dataset and *O1O2O3O4-NSGA-II* for indoor temperature dataset, marking in red the solution selected by the decision-making process.

C. Analysis of the Final Prediction Model

The performance measure $\mathcal{H}$ proposed in this article has been used finally to evaluate the prediction models on datasets R and T. We can highlight the following observations about the best prediction models found in this phase of experimentation.

1) We can estimate the overfitting ratio of a model with the formula $\mathcal{H}_R/\mathcal{H}_T$. A rate greater than 1 indicates that the model is overfitting the training data, and a rate less than 1 indicates that the model is under-fitting. The closer the rate is to 1, the better, since it implies that the results of training and testing are similar. The overfitting ratio of the prediction model found with *O1O2O3* and NSGA-II for the air quality problem is $\mathcal{H}_R/\mathcal{H}_T = 0.6217$, while the overfitting ratio of the prediction model found with *O1O2O3O4* and NSGA-II for the indoor temperature problem is $\mathcal{H}_R/\mathcal{H}_T = 0.5304$. This indicates that the proposed surrogate-assisted approach is not prone to overfitting, since there is no model training during the optimization phase, but rather an evaluation of the surrogate model on the validation dataset V.
2) The prediction model found in this research with *O1O2O3* and NSGA-II for the air quality problem is, not only accurate but also stable, in the sense that the h-steps ahead predictions suffer only a small increase in the prediction error between one step and the next. The prediction model found with *O1O2O3O4* and NSGA-II for the indoor temperature problem is also accurate but suffers from less stability, mainly because the window size has been smaller in this case.

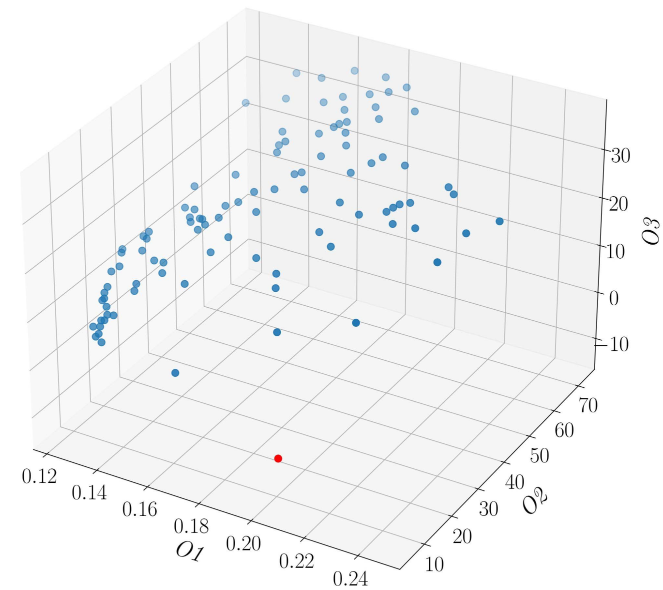

Fig. 3. Pareto front found with *O1O2O3-NSGA-II* (air quality dataset), best run, 100 000 evaluations. Red point represents the model with best average RMSE of seven-steps ahead predictions.

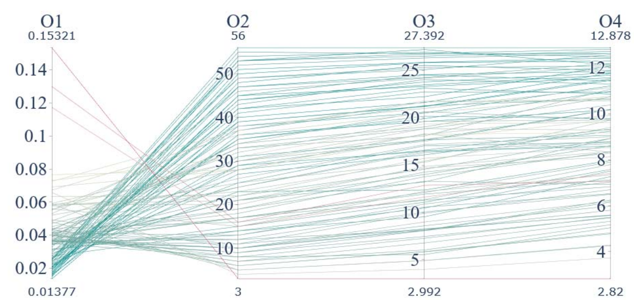

Fig. 4. Parallel coordinates plot of the Pareto front found with *O1O2O3O4-NSGA-II* (indoor temperature dataset), best run, 100 000 evaluations.

D. Comparison With Other FS Approaches

The following analysis can be derived from the comparison with the FS methods described in Section IV-F.

1) None of the compared FS methods beats the *O1O2O3-NSGA-II* method nor the *O1O2O3O4-NSGA-II* method in their respective application problems.
2) The *M1* and *M2* methods based on correlation and reliefF have performed better than the wrapper methods *M3* and *M4* based on linear regression and random forest. The *M1* and *M2* methods, in addition to using the correlation and reliefF filters, respectively, build LSTM neural networks with the subsets of attributes, which

TABLE XII
SELECTED ATTRIBUTES IN AIR QUALITY AN INDOOR TEMPERATURE PROBLEMS

Problem	Method	Selected Attributes
Air quality	*O1O2O3-NSGA-II*	*Lag_NO_1*, *Lag_NO_4*, *Lag_HR_7*, *Lag_NO_X_1*, *Lag_NO_X_2*, *Lag_NO_X_3*, *Lag_NO_X_4*, *Lag_NO_X_6*, *Lag_NO_X_7*, *Lag_DD_2*, *Lag_NO_2_2*, *Lag_NO_2_3*, *Lag_NO_2_4*, *Lag_NO_2_5*, *Lag_NO_2_6*, *Lag_NO_2_7'*
Indoor temperature	*O1O2O3O4-NSGA-II*	*Lag_CO_2_2*, *Lag_W_4*, *Lag_SLS_4*, *Lag_OT_1*, *Lag_OT_2*, *Lag_OT_4*, *Lag_IT_1*, *Lag_IT_2*, *Lag_IT_3*, *Lag_IT_4*

can make a difference with respect to the *M3* and *M4* methods.

3) The embedded method *M5* (based on the cancelOut layer) has not shown good performance, even after adjusting its hyperparameters. The *M5* method is the only FS method that has not selected any attributes.
4) The embedded method *M6* (based on random forest) selects few attributes but performs worse than the best FS methods found in this article.
5) With regard to run times, the FS method *O1O2O3-NSGA-II* found in this work, which has evaluated 100 000 subsets of attributes (using the surrogate model) and has built a maximum of 100 LSTM neural networks (maximum number of non-dominated solutions in the last population), consumes less computational time than the *M4* wrapper method based on random forest, which builds and evaluates 100 000 prediction models. However, the *O1O2O3O4-NSGA-II* method has taken longer than the *M4* method due to the presence of four objective functions. The *M3* wrapper method based on linear regression, although it also builds 100 000 prediction models, consumes a similar run time to the *M1* and *M2* methods, since the linear regression learning algorithm is a really fast method. The embedded methods *M5* and *M6* based on cancelOut and random forest, respectively are the least time-consuming methods. The *M5* method consumes similar to the time required to train an LSTM neural network with all 84 attributes.

E. Interpretation of the Selected Attribute Subsets

In this section, we show the features that each model has selected for forecasting and discuss whether they really are important features from a domain perspective. This analysis strengthens the interpretability aspects of the models. Table XII shows the attributes selected by the *O1O2O3-NSGA-II* and *O1O2O3O4-NSGA-II* methods in the air quality and indoor temperature problems, respectively.

In the air quality problem, lagged variables of the attributes *NO*, NO_X, *DD*, and *HR* were selected. The attributes *NO* and NO_X are closely related to the attribute NO_2 in their chemical composition, and therefore there is a high (Pearson's) correlation between these attributes and the target attribute NO_2. Climate also influences NO_2 concentrations, according to a study by the Leibniz Institute for Tropospheric Research (Germany) [66] commissioned by the State Office for Environment, Agriculture and Geology (LfULG). In this study, it was shown that wind speed and the height of the lowest air layer are the most important factors that determine how much pollutants can accumulate locally. Moreover, it has long been known that weak winds can cause high concentrations of pollutants. The study also showed that high humidity can also reduce the concentration of NO_2, which could be due to the fact that the pollutants deposit more strongly on moist surfaces. The FS method has selected the first six lagged variables (out of 7), which may imply a low incidence in the prediction of the last day of the established time window.

In the indoor temperature problem, lagged variables of the attributes CO_2, *W*, *SLS*, and *OT* have been selected. The presence of a high concentration of CO_2 generated by the occupants of the room through breathing is a consequence of a poor ventilation system that leads to an increase in indoor temperature. On the other hand, it is evident that the external wind, the sunlight on the south facade (in areas of the northern hemisphere, as is the case) and the external temperature are factors that affect the indoor temperature of a room. With respect to the lagged variables of the target, all of them have been selected, which indicates a dependency of the entire time window on the model forecast.

F. Summary of Analysis Results

In summary, the results of the analysis are as follows.

1) All the objectives *O1*, *O2*, *O3*, and *O4* proposed in this research for multiobjective evolutionary FS have proven to be important in the identification of good forecasting models in the cases under study. Particularly, the combinations *O1O2O3* and *O1O2O3O4* have turned out to be the best in the problem of air quality and indoor temperature, respectively.
2) The NSGA-II algorithm has shown the best performance, in terms of hypervolume, than the rest of the MOEAs in solving the multiobjective optimization problems proposed for FS in this article.
3) The prediction models found in this study with the proposed techniques present low overfitting rates and are stable with respect to multistep ahead predictions.
4) The proposed FS methods outperform other powerful filter-wrapper FS methods based on correlation, relief, and LSTM, wrapper FS methods based on linear regression and random forest, and embedded FS methods based on cancelOut and random forest.
5) The models found in this study for air quality prediction and indoor temperature prediction contain a relatively low number of attributes that are easy to interpret in the problem domain.

VI. CONCLUSIONS AND FUTURE WORK

In this article, we have proposed new FS methods that are particularly appropriate for building prediction models that require a high computational training cost, such as models based on deep learning. The proposed methods search for subsets of attributes by means of a multiobjective evolutionary strategy and using an LSTM neural network which acts as a surrogate model to evaluate candidate subsets of attributes, together with correlation and reliefF filters and cardinality minimization of attribute subsets. In the experimentation process, different optimization models of two, three, and four objectives have been tested, which combine the evaluation by means of the surrogate model with the correlation and reliefF filters. Different state-of-the-art multiobjective and many-objective evolutionary algorithms have also been tested, and the algorithm NSGA-II has shown the best performance. Two datasets have been used in the experiments, the first with air quality time series data in south-eastern Spain with data measured daily for 4 years, and the second with time series data related to the indoor temperature of a home domotic house. The data was divided into train, validation, and test sets. To measure the performance of the prediction models, a performance measure multicriteria metric has been proposed, which takes into account the average RMSE, MAE, and CC in a multistep ahead prediction horizon. In the air quality forecasting problem, the best FS method found was *O1O2O3-NSGA-II*, selecting 16 attributes out of a total of 84 and considerably improving the predictive capacity of the LSTM neural network with the complete set of attributes. In the indoor temperature forecasting problem, the best FS method found was *O1O2O3O4-NSGA-II*, which selected 10 attributes out of a total of 57, also improving the performance of the LSTM with all the attributes. They also show stability in the predictions without the presence of overfitting. The investigated FS methods *O1O2O3-NSGA-II* and *O1O2O3O4-NSGA-II* outperform, in their respective application scenario, FS methods of the wrapper, hybrid filter-wrapper, and embedded types. Finally, we have verified the importance of the selected attributes with the *O1O2O3-NSGA-II* and *O1O2O3O4-NSGA-II* methods in the expert context.

Future works will include the use of this novel approach with other deep learning algorithms such as GRU and CNN as well as other multistep ahead forecasting strategies. In addition, predictions can be made for other harmful chemical compounds such as NO_X, $PM_{2.5}$, PM_{10}, and so on. Besides, this approach can be applied to classification problems and imaging processing [67]. We are currently working on a new many-objective evolutionary algorithm based on decomposition, performance indicators, and reference points sampled on the Pareto front, and on the development of a novel multisurrogate-assisted MOEA for FS method in order to improve the generalization error applied to time series forecasting problems. Finally, we are considering the application of the proposed technique in a spatio-temporal forecast scenario.

REFERENCES

[1] M. Kuhn and K. Johnson, *Feature Engineering and Selection: A Practical Approach for Predictive Models*. Boca Raton, FL, USA: CRC Press, 2019.

[2] K. Deb, *Multi-Objective Optimization Using Evolutionary Algorithms*. Hoboken, NJ, USA: Wiley, 2001.

[3] S. Hochreiter and J. J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 80–1735, 1997.

[4] M. Kowalska, M. Skrzypek, M. Kowalski, and J. Cyrys, "Effect of NO_X and NO_2 concentration increase in ambient air to daily bronchitis and asthma exacerbation, silesian voivodeship in Poland," *Int. J. Environ. Res. Public Health*, vol. 17, no. 3, p. 754, Jan. 2020.

[5] R. W. Atkinson, B. K. Butland, H. R. Anderson, and R. L. Maynard, "Long-term concentrations of nitrogen dioxide and mortality: A meta-analysis of cohort studies," *Epidemiology*, vol. 29, no. 4, pp. 460–472, Jul. 2018.

[6] U. Kamath and J. Liu, *Explainable Artificial Intelligence: An Introduction to Interpretable Machine Learning*. Cham, Switzerland: Springer, 2021.

[7] H. Liu and H. Motoda, *Feature Selection for Knowledge Discovery and Data Mining*. Norwell, MA, USA: Kluwer, 1998.

[8] R. Ruiz, S. J. Aguilar-Ruiz, and C. J. Riquelme and N. Díaz-Díaz, "Analysis of feature rankings for classification," in *Advances in Intelligent Data Analysis* (Lecture Notes in Computer Science). Berlin, Germany: Springer, 2005, pp. 362–372.

[9] I. Kononenko, E. Šimec, and M. Robnik-Šikonja, "Overcoming the myopia of inductive learning algorithms with RELIEFF," *Appl. Intell.*, vol. 7, no. 1, pp. 39–55, Jan. 1997.

[10] X. Zhang, M. Fan, D. Wang, P. Zhou, and D. Tao, "Top-k feature selection framework using robust 0–1 integer programming," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 7, pp. 3005–3019, Jul. 2021.

[11] E. Schaffernicht, C. Möller, K. Debes, and H.-M. Gross, "Forward feature selection using residual mutual information," in *Proc. ESANN*, 2009, pp. 583–588.

[12] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach. Learn.*, vol. 46, nos. 1–3, pp. 389–422, 2004.

[13] F. Jiménez, G. Sánchez, J. M. García, G. Sciavicco, and L. Miralles, "Multi-objective evolutionary feature selection for online sales forecasting," *Neurocomputing*, vol. 234, pp. 75–92, Apr. 2017.

[14] M. Hall, "Correlation-based feature selection for machine learning," *Dept. Comput. Sci.*, vol. 19, Jun. 2000.

[15] M. Dash and H. Liu, "Consistency-based search in feature selection," *Artif. Intell.*, vol. 151, nos. 1–2, pp. 155–176, Dec. 2003.

[16] L. Yu and H. Liu, "Redundancy based feature selection for microarray data," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, New York, NY, USA, 2004, pp. 737–742.

[17] S. Lei, "A feature selection method based on information gain and genetic algorithm," in *Proc. Int. Conf. Comput. Sci. Electron. Eng.*, vol. 2, Mar. 2012, pp. 355–358.

[18] K. Pearson, "X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling," *Philos. Mag.*, vol. 50, no. 320, pp. 157–175, 1900.

[19] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Soc., B, Methodol.*, vol. 58, no. 1, pp. 267–288, 1994.

[20] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann, 1993.

[21] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification Regression Trees*. Boca Raton, FL, USA: CRC Press, 1984.

[22] V. Bolón-Canedo and A. Alonso-Betanzos, "Recent advances in ensembles for feature selection," in *Intelligent Systems Reference Library*. Cham, Switzerland: Springer, 2018, pp. 13–37.

[23] V. Borisov, J. Haug, and G. Kasneci, "Cancelout: A layer for feature selection in deep neural networks," in *Artificial Neural Networks and Machine Learning*, I. V. Tetko, V. Kůrková, P. Karpov, and F. Theis, Eds. Cham, Switzerland: Springer, 2019, pp. 72–83.

[24] Y. Huang, W. Jin, Z. Yu, and B. Li, "Supervised feature selection through deep neural networks with pairwise connected structure," *Knowl.-Based Syst.*, vol. 204, Sep. 2020, Art. no. 106202.

[25] L. Zhou et al., "Wheat kernel variety identification based on a large near-infrared spectral dataset and a novel deep learning-based feature selection method," *Frontiers Plant Sci.*, vol. 11, p. 1682, Nov. 2020.

[26] L. Cun et al., "Handwritten digit recognition with a back-propagation network," in *Advances in Neural Information Processing Systems*. San Mateo, CA, USA: Morgan Kaufmann, 1990, pp. 396–404.

[27] L. Wang, X. Yan, M.-L. Liu, K.-J. Song, X.-F. Sun, and W.-W. Pan, "Prediction of RNA-protein interactions by combining deep convolutional neural network with feature selection ensemble method," *J. Theor. Biol.*, vol. 461, pp. 230–238, Jan. 2019.

[28] X.-F. Yuan, L. Li, Y. Shardt, Y.-L. Wang, and C.-H. Yang, "Deep learning with spatiotemporal attention-based LSTM for industrial soft sensor model development," *IEEE Trans. Ind. Electron.*, vol. 68, no. 5, pp. 4404–4414, May 2020.

[29] J. Wang, H. Zhang, J. Wang, Y. Pu, and N. R. Pal, "Feature selection using a neural network with group lasso regularization and controlled redundancy," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 3, pp. 1110–1123, Mar. 2021.

[30] W. Zheng, S. Chen, Z. Fu, F. Zhu, H. Yan, and J. Yang, "Feature selection boosted by unselected features," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 9, pp. 4562–4574, Sep. 2022.

[31] X. Wu, X. Xu, J. Liu, H. Wang, B. Hu, and F. Nie, "Supervised feature selection with orthogonal regression and feature weighting," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 5, pp. 1831–1838, May 2021.

[32] P. Zhou, P. Li, S. Zhao, and X. Wu, "Feature interaction for streaming feature selection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 10, pp. 4691–4702, Oct. 2021.

[33] H. Ishibuchi et al., "Evolutionary multi-criterion optimization," in *Proc. 11th Int. Conf.*, Shenzhen, China: Springer, vol. 12654, Mar. 2021, pp. 1–781.

[34] Q. Al-Tashi et al., "Binary multi-objective grey wolf optimizer for feature selection in classification," *IEEE Access*, vol. 8, pp. 106247–106263, 2020.

[35] T. Niu, J. Wang, H. Lu, W. Yang, and P. Du, "Developing a deep learning framework with two-stage feature selection for multivariate financial time series forecasting," *Expert Syst. Appl.*, vol. 148, Jun. 2020, Art. no. 113237.

[36] L. Shu, F. He, X. Hu, and H. Li, "A novel feature selection with many-objective optimization and learning mechanism," in *Proc. IEEE 24th Int. Conf. Comput. Supported Cooperat. Work Design (CSCWD)*, May 2021, pp. 684–689.

[37] K. Deb and H. Jain, "An evolutionary many-objective optimization algorithm using reference-point-based nondominated sorting approach, Part I: Solving problems with box constraints," *IEEE Trans. Evol. Comput.*, vol. 18, no. 4, pp. 577–601, Aug. 2014.

[38] Y. Jin, "Surrogate-assisted evolutionary computation: Recent advances and future challenges," *Swarm Evol. Comput.*, vol. 1, no. 2, pp. 61–70, Jun. 2011.

[39] Z. Lv, L. Wang, Z. Han, J. Zhao, and W. Wang, "Surrogate-assisted particle swarm optimization algorithm with Pareto active learning for expensive multi-objective optimization," *IEEE/CAA J. Autom. Sinica*, vol. 6, no. 3, pp. 838–849, May 2019.

[40] T. H. Nguyen, D. Nong, and K. Paustian, "Surrogate-based multi-objective optimization of management options for agricultural landscapes using artificial neural networks," *Ecol. Model.*, vol. 400, pp. 1–13, May 2019.

[41] Z. Jiang, Y. Zhang, and J. Wang, "A multi-surrogate-assisted dual-layer ensemble feature selection algorithm," *Appl. Soft Comput.*, vol. 110, Oct. 2021, Art. no. 107625.

[42] H. Liu, Z. Duan, and C. Chen, "A hybrid multi-resolution multi-objective ensemble model and its application for forecasting of daily $PM_{2.5}$ concentrations," *Inf. Sci.*, vol. 516, pp. 266–292, 2020.

[43] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Trans. Evol. Comput.*, vol. 6, no. 2, pp. 182–197, Apr. 2002.

[44] S. Zhang, Y. Chen, W. Zhang, and R. Feng, "A novel ensemble deep learning model with dynamic error correction and multi-objective ensemble pruning for time series forecasting," *Inf. Sci.*, vol. 544, pp. 427–445, Jan. 2021.

[45] J. Wang, H. Li, H. Yang, and Y. Wang, "Intelligent multivariable air-quality forecasting system based on feature selection and modified evolving interval type-2 quantum fuzzy neural network," *Environ. Pollut.*, vol. 274, Apr. 2021, Art. no. 116429.

[46] P. Du, J. Wang, Y. Hao, T. Niu, and W. Yang, "A novel hybrid model based on multi-objective Harris hawks optimization algorithm for daily $PM_{2.5}$ and PM_{10} forecasting," *Appl. Soft Comput.*, vol. 96, Nov. 2020, Art. no. 106620.

[47] F. Jiménez, J. Palma, G. Sánchez, D. Marín, M. D. Francisco Palacios, and M. D. Lucía López, "Feature selection based multivariate time series forecasting: An application to antibiotic resistance outbreaks prediction," *Artif. Intell. Med.*, vol. 104, Apr. 2020, Art. no. 101818.

[48] A. González-Vidal, F. Jiménez, and A. F. Gómez-Skarmeta, "A methodology for energy multivariate time series forecasting in smart buildings based on feature selection," *Energy Buildings*, vol. 196, pp. 71–82, Aug. 2019.

[49] R. Espinosa, J. Palma, F. Jiménez, J. Kamińska, G. Sciavicco, and E. Lucena-Sánchez, "A time series forecasting based multi-criteria methodology for air quality prediction," *Appl. Soft Comput.*, vol. 113, Dec. 2021, Art. no. 107850.

[50] R. Espinosa, F. Jiménez, and J. Palma, "Multi-objective evolutionary spatio–temporal forecasting of air pollution," *Future Gener. Comput. Syst.*, vol. 136, pp. 15–33, Nov. 2022.

[51] B. Xue, M. Zhang, W. N. Browne, and X. Yao, "A survey on evolutionary computation approaches to feature selection," *IEEE Trans. Evol. Comput.*, vol. 20, no. 4, pp. 606–626, Aug. 2016.

[52] L. J. Eshelman, "The CHC adaptive search algorithm: How to have safe search when engaging in nontraditional genetic recombination," *Found. Genetic Algorithms*, vol. 1, pp. 265–283, Jan. 1991.

[53] L. Davis, *Handbook of Genetic Algorithms*. New York, NY, USA: Van Nostrand Reinhold, 1991.

[54] G. Bontempi, S. B. Taieb, and Y. Borgne, "Machine learning strategies for time series forecasting," in *Proc. eBISS*, 2013, pp. 62–77.

[55] S. B. Taieb and A. F. Atiya, "A bias and variance analysis for multistep-ahead time series forecasting," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 1, pp. 62–76, Jan. 2016.

[56] R. Wen, K. Torkkola, B. Narayanaswamy, and D. Madeka, "A multi-horizon quantile recurrent forecaster," 2017, *arXiv:1711.11053*.

[57] J. Brownlee. (Dec. 2020). *Time Series Forecasting as Supervised Learning*. [Online]. Available: https://machinelearningmastery.com/time-series-forecasting-supervised-learning/

[58] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. ICML*, 2010, pp. 1–8.

[59] Q. Zhang and H. Li, "MOEA/D: A multiobjective evolutionary algorithm based on decomposition," *IEEE Trans. Evol. Comput.*, vol. 11, no. 6, pp. 712–731, Dec. 2007.

[60] E. Zitzler and S. Künzli, "Indicator-based selection in multiobjective search," in *Parallel Problem Solving From Nature*, X. Yao et al., Eds. Berlin, Germany: Springer, 2004, pp. 832–842.

[61] Z. Fan et al., "An improved epsilon constraint-handling method in MOEA/D for CMOPs with large infeasible regions," *Soft Comput.*, vol. 23, no. 23, pp. 12491–12510, 2019.

[62] E. Zitzler, M. Laumanns, and L. Thiele, "SPEA2: Improving the strength Pareto evolutionary algorithm," Comput. Eng. Netw. Lab., Institut für Technische Informatik und Kommunikationsnetze, ETH Zurich, Zürich, Switzerland, Tech. Rep. 103, 2001, vol. 103.

[63] J. B. Kollat and P. M. Reed, "The value of online adaptive search: A performance comparison of nsgaii, ϵ-nsgaii and ϵmoea," in *Evolutionary Multi-Criterion Optimization*, C. A. C. Coello, A. H. Aguirre, and E. Zitzler, Eds. Berlin, Germany: Springer, 2005, pp. 386–398.

[64] (2017). Dua, Dheeru and Graff, Casey. *UCI Machine Learning Repository*. [Online]. Available: http://archive.ics.uci.edu/ml

[65] F. Zamora-Martínez, P. Romeu, P. Botella-Rocamora, and J. Pardo, "On-line learning of indoor temperature forecasting models towards energy efficiency," *Energy Buildings*, vol. 83, pp. 162–172, Nov. 2014.

[66] D. V. Pinxteren, S. Düsing, and A. W. A. Herrmann, "Meteorological influences on nitrogen dioxide: Influence of weather conditions and weathering on nitrogen dioxide concentrations in outdoor air 2015 to 2018," Landesamt für Umwelt, Landwirtschaft und Geologie, Dresden, Germany, Tech. Rep. 2, 2020.

[67] A. Glowacz, "Fault diagnosis of electric impact drills using thermal imaging," *Measurement*, vol. 171, Feb. 2021, Art. no. 108815.