UNIVERSIDAD DE MURCIA

ESCUELA INTERNACIONAL DE DOCTORADO

TESIS DOCTORAL

Multi-objective evolutionary feature selection with deep learning applied to air quality spatio-temporal forecasting

Selección evolutiva multi-objetivo de atributos para deep learning aplicada al pronóstico de series espacio-temporales de calidad del aire

D.ª Raquel Espinosa Fernández

2023

UNIVERSIDAD DE MURCIA

ESCUELA INTERNACIONAL DE DOCTORADO

TESIS DOCTORAL

Multi-objective evolutionary feature selection with deep learning applied to air quality spatio-temporal forecasting

Selección evolutiva multi-objetivo de atributos para deep learning aplicada al pronóstico de series espacio-temporales de calidad del aire

Autor: D.ª Raquel Espinosa Fernández

Director/es: D. Fernando Jiménez Barrionuevo
D. José Tomás Palma Méndez

UNIVERSIDAD DE
MURCIA

DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD DE LA TESIS PRESENTADA EN MODALIDAD DE COMPENDIO O ARTÍCULOS PARA OBTENER EL TÍTULO DE DOCTOR

Aprobado por la Comisión General de Doctorado el 19-10-2022

D./Dña. Raquel Espinosa Fernández

doctorando del Programa de Doctorado en

Informática

de la Escuela Internacional de Doctorado de la Universidad Murcia, como autor/a de la tesis presentada para la obtención del título de Doctor y titulada:

Multi-objective evolutionary feature selection with deep learning applied to air quality spatio-temporal forecasting / Selección evolutiva multi-objetivo de atributos para deep learning aplicada al pronóstico de series espacio-temporales de calidad del aire

y dirigida por,

D./Dña. Fernando Jiménez Barrionuevo

D./Dña. José Tomás Palma

D./Dña.

DECLARO QUE:

La tesis es una obra original que no infringe los derechos de propiedad intelectual ni los derechos de propiedad industrial u otros, de acuerdo con el ordenamiento jurídico vigente, en particular, la Ley de Propiedad Intelectual (R.D. legislativo 1/1996, de 12 de abril, por el que se aprueba el texto refundido de la Ley de Propiedad Intelectual, modificado por la Ley 2/2019, de 1 de marzo, regularizando, aclarando y armonizando las disposiciones legales vigentes sobre la materia), en particular, las disposiciones referidas al derecho de cita, cuando se han utilizado sus resultados o publicaciones.

Además, al haber sido autorizada como compendio de publicaciones o, tal y como prevé el artículo 29.8 del reglamento, cuenta con:

- *La aceptación por escrito de los coautores de las publicaciones de que el doctorando las presente como parte de la tesis.*
- *En su caso, la renuncia por escrito de los coautores no doctores de dichos trabajos a presentarlos como parte de otras tesis doctorales en la Universidad de Murcia o en cualquier otra universidad.*

Del mismo modo, asumo ante la Universidad cualquier responsabilidad que pudiera derivarse de la autoría o falta de originalidad del contenido de la tesis presentada, en caso de plagio, de conformidad con el ordenamiento jurídico vigente.

En Murcia, a 20 de Febrero de 2023

Fdo.: Raquel Espinosa Fernández

Esta DECLARACIÓN DE AUTORIA Y ORIGINALIDAD debe ser insertada en la primera página de la tesis presentada para la obtención del título de Doctor.

Agradecimientos

Me gustaría agradecer a todas aquellas personas que me han ayudado y apoyado durante el desarrollo de esta tesis doctoral, y sin las cuales no habría sido posible.

En primer lugar, agradecer a mis directores Fernando Jiménez y José Palma su inestimable dedicación, orientación y sabios consejos a lo largo de estos años de tesis. Vuestras críticas, conocimientos y experiencia me han alentado durante mi investigación y me han ayudado a dar forma a esta tesis. También hacer una mención especial a mi supervisora Josiane Parreira de mi estancia en Siemens AG Österreich por acogerme y por nuestros debates tan interesantes y enriquecedores.

A mis padres, Jesús y María Ginesa, por su constante apoyo y cariño durante toda mi etapa universitaria. Ellos mejor que nadie saben que convivir con una doctoranda no siempre es fácil. Tampoco olvidar a mis tías Toty, Cruz y Ana quienes siempre han tenido palabras de ánimo para mí.

A mis amigos y compañeros de aventuras, Mari, Hamisa, Antolín y Ana por todos los buenos momentos que me habéis dado y por descubrirme lo que son los amigos de verdad.

A Jesús, gracias por tu paciencia y entrega para conmigo, nunca podré agradecerte lo suficiente tanta ayuda y el haberme enseñado a ver el lado bueno de las cosas.

A los que se fueron, por cuidarme y quererme allá donde estén.

Contents

List of Figures

List of Tables

List of publications

This Doctoral Thesis is presented as a compendium of the following publications, being the PhD student the main author in all of them:

- R. Espinosa, J. Palma, F. Jiménez, J. Kamińska, G. Sciavicco, E. Lucena-Sánchez, A time series forecasting based multi-criteria methodology for air quality prediction, Applied Soft Computing 113 (2021) 107850.
- R. Espinosa, F. Jiménez, J. Palma, Multi-objective evolutionary spatio-temporal forecasting of air pollution, Future Generation Computer Systems 136 (2022) 15–33.
- R. Espinosa, F. Jiménez, J. Palma, Multi-surrogate assisted multi-objective evolutionary algorithms for feature selection in regression and classification problems with time series data, Information Sciences 622 (2023) 1064–1091.
- R. Espinosa, F. Jiménez, J. Palma, Surrogate-assisted and filter-based multi-objective evolutionary feature selection for deep learning, IEEE Transactions on Neural Networks and Learning Systems (2023) 1–15.

Resumen

El aumento en los últimos años de la cantidad de información disponible ha hecho que cada vez se desarrollen modelos predictivos con una mayor dimensionalidad y complejidad. Esto puede mitigarse gracias a un proceso de *selección de atributos*, el cual permite reducir la dimensionalidad de los datos de entrada a la hora de construir modelos predictivos, eliminando atributos redundantes y/o irrelevantes para la tarea que se está aplicando. Los tres tipos principales de métodos para la selección de atributos son: *filter*, *wrapper* y *embedded*. Los métodos filter separan el proceso de selección de atributos del algoritmo de aprendizaje, por lo que la influencia de este último no interactúa en la selección. Normalmente son métodos estadísticos. Los métodos wrapper usan la precisión predictiva de un algoritmo de aprendizaje predeterminado para establecer la calidad de los atributos seleccionados. Los métodos embedded logran el ajuste del modelo y la selección de características simultáneamente. Tanto los filters como los embeddeds son métodos computacionalmente rápidos. Sin embargo, los métodos wrapper son los que mejores resultados obtienen pese a ser muy costosos computacionalmente, especialmente si se aplican junto con modelos de *deep learning* y en escenarios con muchos atributos, como puede ser el caso de algunas aplicaciones basadas en *series temporales* de datos. En esta tesis doctoral se ha formalizado la selección de atributos como un problema de optimización multi-objetivo, lo cual ha permitido la aplicación de unas estrategias de búsqueda de la solución óptima conocidas como *algoritmos evolutivos multi-objetivo* (MOEAs). Los problemas de optimización multi-objetivo se caracterizan por tener funciones objetivo conflictivas. Su principal propósito es encontrar un balance óptimo entre todas las funciones objetivo. Este equilibrio establece un conjunto de soluciones denominadas *no-dominadas*, las cuales constituyen el *frente de Pareto*. El uso de *modelos sustitutos*, también conocidos como meta-modelos, permite aproximar la función fitness del algoritmo evolutivo reduciendo así el tiempo computacional respecto al consumido por los métodos wrapper de selección de atributos convencionales basados en deep learning.

En el transcurso de esta tesis se ha desarrollado una metodología genérica y nuevas métricas para la comparación de modelos de *machine learning* y deep learning, estableciendo así el mejor modelo para usar en un problema determinado de series temporales. La metodología permite, además, establecer distintos *tamaños de ventana* para la aplicación del método de la *ventana deslizante*, pudiendo encontrar así el tamaño más apropiado para hacer predicciones confiables y robustas. Las fases de la metodología son las siguientes: transformación por ventana deslizante con diferentes tamaños de ventana, ajuste de hiper-parámetros, test estadísticos, toma de decisiones multi-criterio considerando la *raíz del error cuadrático medio* (RMSE), el *error medio absoluto* (MAE) y el *coeficiente de*

correlación (CC) y, por último, generación de pasos adelante. Los algoritmos de aprendizaje usados para validar esta metodología han sido los siguientes: *redes neuronales convolucionales de 1 dimensión* (1D-CNN), *gated recurrent unit* (GRU), *long short-term memory* (LSTM), *árboles aleatorios* (RF), *lasso* y *máquinas de soporte vectorial* (SVM) con kernel radial. El proceso de toma de decisiones multi-criterio tiene en cuenta la exactitud y robustez del RMSE, MAE y CC de los modelos, y reúne estos criterios en una única métrica ponderada denominada *goodness*. El modelo con menor *goodness* es el finalmente seleccionado.

Se ha propuesto una nueva técnica fundamentada en las propiedades espacio-temporales de los datos para inferir información de zonas de las que no se tienen datos. Para ello, se ha formalizado el problema de predicción de calidad del aire como un problema de optimización multi-objetivo. Así, cada objetivo es el RMSE de un modelo predictivo basado en *regresión lineal* (LR) con datos de una estación de monitoreo. Estos objetivos se minimizan para conseguir las mejores predicciones en cada ubicación. En el contexto de aplicación utilizado en esta tesis, se trata de un problema de optimización de 3 objetivos. Los algoritmos evolutivos multi-objetivo evaluados han sido: *NSGA-II*, *MOEA/D* y *SPEA2*. Los frentes de Pareto resultantes del algoritmo evolutivo son la entrada para construir un modelo de aprendizaje *ensemble*. El método de aprendizaje ensemble se basa en la técnica de *stacking*. Se han utilizado los siguientes algoritmos de aprendizaje para entrenar el modelo ensemble: RF, LR, SVM, *redes neuronales cuasi-recurrentes* (QRNN), *perceptrón multicapa* (MLP), *k vecinos más cercanos* (KNN) y *ZeroR*. Con el conjunto entrenado, se realiza el proceso de pronóstico en varios pasos adelante para hacer las predicciones con los datos de test de la estación de monitoreo, los cuales no han sido vistos por el algoritmo evolutivo.

Por otra parte, se han formalizado problemas de optimización con hasta cuatro objetivos, basados en métodos filter, wrapper e híbridos, para realizar la selección de atributos. Los algoritmos de *correlación* y *reliefF* se han adaptado para la evaluación de subconjuntos de atributos y se han utilizado para definir los objetivos relacionados con los métodos filter. Gracias a un modelo sustituto, este método permite el uso de modelos de deep learning como algoritmo de aprendizaje de un método wrapper pero sin el inconveniente del elevado coste computacional que ello conlleva. La idea subyacente al planteamiento propuesto es la siguiente: el modelo sustituto se evalúa con un conjunto de datos de test consistente en el conjunto de datos de test original en el que los valores de los atributos no seleccionados se fijan en un valor constante $\alpha = 0$ en todas las muestras. Cuando un atributo es redundante o irrelevante, la evaluación del modelo sustituto, que ha sido entrenado con todos los atributos, se ve poco afectada por este cambio en el conjunto de datos de prueba. Sin embargo, cuando un atributo es relevante, la evaluación del modelo sustituto se verá muy alterada, ya que el atributo ha tenido una gran influencia en el entrenamiento del modelo sustituto. Se ha estudiado cuál es el mejor algoritmo evolutivo multi-objetivo entre los algoritmos NSGA-II, *NSGA-III*, MOEA/D, SPEA2, *IBEA*, ϵ-*NSGA-II* y ϵ-*MOEA*, y el rendimiento de las predicciones con un modelo sustituto basado en una red neuronal LSTM. Adicionalmente, se propone una nueva métrica multi-criterio de rendimiento, $\mathcal{H}$, la cual permite ajustar la importancia de las métricas que la forman, pudiendo dar más peso a la que sea más importante para el problema tratado, pero sin perder la información adicional que aportan el resto de métricas. Facilita así la comparación de diferentes modelos y sus predicciones a h-pasos adelante. La métrica es independiente del horizonte de predicción por lo que se puede utilizar la más adecuada para cada problema. Además, permite obtener modelos de predicción más robustos, ya que considera varias métricas y las agrega. También se ha evaluado el comportamiento del uso de múltiples modelos sustitutos para lograr una mejor capacidad de generalización de los modelos predictivos tanto en problemas de regresión como de clasificación.

En los enfoques descritos anteriormente, el modelo sustituto mantiene siempre la misma información que al principio del método. Esto, en cierto modo, es una desventaja, ya que no se tiene en cuenta la información subyacente obtenida durante el transcurso del algoritmo evolutivo. Por este motivo, se han propuesto dos enfoques para actualizar el modelo sustituto, uno basado en el *aprendizaje incremental* y otro basado en la actualización de la base de datos y la construcción de un nuevo modelo sustituto. El modelo sustituto se obtiene construyendo "offline" un meta-modelo de aprendizaje a partir de un conjunto de muestras de selecciones de atributos y sus evaluaciones reales se obtienen entrenando una red neuronal LSTM para cada muestra. Por tanto, en esta tesis se han propuesto hasta un total de cuatro nuevos métodos evolutivos multi-objetivo asistidos por sustitutos, de los cuales dos utilizan enfoques sin control de evolución, y los otros dos son técnicas con control de evolución fijo basado en generación.

Durante la estancia de doctorado en Siemens AG Österreich en Viena (Austria), se ha llevado a cabo un estudio de diferentes técnicas de clasificación y *clustering* de series temporales. Estos experimentos han dado lugar a una mejor compresión de las series temporales aplicadas en un contexto de clasificación. La experiencia adquirida ha sido útil para el desarrollo de un método aplicable a problemas de regresión y clasificación.

Todos estos experimentos se han realizado sobre diversos conjuntos de datos de series temporales. El primer conjunto de datos contiene información de la calidad del aire entre 2015 y 2017 de una estación de monitorización situada en la ciudad de Wroclaw (Polonia). También se han usado cuatro conjuntos de datos de calidad del aire entre 2017 y 2020 de cuatro estaciones situadas en La Aljorra, Alcantarilla, Lorca y Valle de Escombreras, todas ellas dentro de la Región de Murcia (España). Estos datos se han extraído del portal de calidad del aire de la Región de Murcia. Otro conjunto de datos usado contiene datos procedentes de una casa domótica en Valencia (España) tomados entre marzo y mayo de 2012. Este conjunto de datos se ha extraído del *UCI Machine Learning Repository*. Por último, el conjunto de datos privado usado en la estancia de doctorado contiene mediciones horarias de energía, calefacción y agua procedentes de varios sensores situados en el interior de edificios inteligentes. Todos los resultados han sido validados con diversos test estadísticos, incluyendo el test de *Diebold-Mariano*, específico para series temporales, y se han comparado con otros métodos del estado del arte, del tipo filter, wrapper y embedded. La implementación de los distintos métodos presentados se ha llevado a cabo en *Python*. Para la implementación de los algoritmos evolutivos multi-objetivo se ha empleado la librería *Platypus*.

Las principales conclusiones derivadas de esta tesis tras el desarrollo y ejecución de todos los experimentos planteados son las siguientes:

- La adopción de una metodología completa para la evaluación y comparación de algoritmos de aprendizaje ha permitido obtener resultados unificados y adaptados para resolver cualquier problema de predicción con series temporales.

- Las redes neuronales recurrentes, como LSTM y GRU, han sido capaces de captar la complejidad de las series temporales y construir modelos predictivos precisos y fiables. Entre las técnicas de machine learning analizadas, RF ha presentado un rendimiento satisfactorio cuando se aplica a la predicción de series temporales.

- Un proceso de toma de decisiones multi-criterio ha permitido agrupar varias métricas de rendimiento y establecer una comparación más adecuada entre diferentes algoritmos de aprendizaje en el contexto de problemas de pronóstico de series temporales.

- Para el pronóstico de la calidad del aire con series temporales en una zona de la que no se dispone de información, la predicción se ha aproximado con algoritmos evolutivos multi-objetivo utilizando las previsiones de otras zonas geográficamente cercanas.

- Los algoritmos evolutivos multi-objetivo asistidos por sustitutos han permitido la selección de atributos en problemas costosos como la predicción de series temporales basada en deep learning. Además, la reducción de dimensionalidad ha simplificado los modelos predictivos construidos aumentando así su interpretabilidad y, ayudando a dejar de percibir los modelos como una “caja negra”.

- El uso de un MOEA asistido por sustitutos con un algoritmo de aprendizaje profundo para la selección de atributos ha conseguido encontrar un subconjunto satisfactorio de atributos en un tiempo computacional más corto en comparación con un método de selección de atributos de tipo wrapper convencional.

- Entre todos los MOEAs estudiados, NSGA-II es el que mejores resultados ha obtenido en términos de hipervolumen, en comparación con otros MOEAs del estado del arte.

- El enfoque de control de la evolución fija basado en la generación de elementos permite añadir información de forma eficiente a los modelos sustitutos dentro del proceso de selección de atributos. Aunque de esta forma se encuentran mejores subconjuntos de atributos y se mejoran los resultados de la predicción, se hace a costa de aumentar el tiempo de cómputo del proceso, ya que después de un número fijo de evaluaciones del algoritmo evolutivo hay que volver a entrenar y/o incrementar el modelo sustituto. No obstante, es más eficiente que los métodos convencionales de selección de características basados en wrappers.

- Se han identificado modelos de predicción en diversos contextos reales (Polonia, Murcia, Valencia) que potencialmente permiten realizar previsiones en un futuro próximo y que pueden ayudar a las instituciones a tomar decisiones en materia medioambiental. Otro factor significativo de los métodos propuestos en el ámbito medioambiental es que, al tener un menor tiempo computacional, se puede reducir la huella de carbono, contribuyendo al *Pacto Verde Europeo*. En el ámbito social, los métodos propuestos contribuyen a la *Inteligencia Artificial Explicable*, y se alinean con los objetivos de iniciativas como el *Libro Blanco en Inteligencia Artificial* de la Comisión Europea o la *Estrategia Nacional de Inteligencia Artificial* de España.

Tras el estudio descrito en esta tesis, se plantean las siguientes líneas de investigación a desarrollar en el futuro:

- Incluir el proceso de selección de características dentro del enfoque espacio-temporal con LR, añadiendo un nuevo objetivo dentro del problema de optimización que trate de minimizar el número de atributos seleccionados en LR. También se estudiará esta aplicación en otros algoritmos de aprendizaje más complejos, como las redes neuronales.

- Aplicar los métodos de selección de atributos del algoritmo evolutivo multi-objetivo asistido por sustitutos y asistido por múltiples sustitutos para la predicción de otras series temporales relacionadas con la calidad del aire como CO_2, $PM_{2,5}$ o PM_{10} y comparar su rendimiento con los resultados de NO_2.

- Utilizar otros algoritmos de aprendizaje profundo como GRU o QRNN como modelo sustituto dentro del método evolutivo multi-objetivo de selección de características y comparar su rendimiento con el método actual con LSTM. Además, se analizará el uso de otras estrategias de predicción de múltiples pasos adelante, como la previsión multipaso directa o la híbrida directa-recursiva.

- Los MOEAs se han utilizado con éxito en la búsqueda de la arquitectura óptima en modelos predictivos, especialmente en aprendizaje profundo. En el desarrollo de esta tesis se ha demostrado que el uso de MOEAs para la selección de atributos con un modelo basado en LSTM obtiene buenos resultados. Sin embargo, no existen trabajos en la actualidad que combinen la búsqueda de arquitecturas con la selección de atributos, por lo que es un campo de estudio muy interesante y que será abordado en futuras investigaciones, tanto para problemas de regresión como de clasificación.

- Aplicar los métodos de selección de atributos propuestos en otros campos como el reconocimiento de imágenes y el procesamiento del lenguaje natural.

Abstract

The increase in the amount of information available in recent years has led to the development of predictive models with high dimensionality and complexity. This can be mitigated using a *feature selection* process, which allows the reduction of the dimensionality of the input data when building predictive models, thus removing redundant and/or irrelevant features for a particular task. One of the most relevant and best-performing types of feature selection are *wrapper* methods. However, they are computationally expensive, especially if applied together with *deep learning* models and in scenarios with many attributes, as may be the case for some applications based on *time series* data. In this PhD thesis, feature selection has been formalized as a multi-objective optimization problem, which allows the application of optimal solution search strategies known as *multi-objective evolutionary algorithms*. In addition, the use of a *surrogate-assisted models* lets the approximation of the fitness function of the evolutionary algorithm thus reducing the computational time compared to the consumed by a conventional deep learning-based wrapper.

In the course of this thesis, a methodology and new error metrics have been developed for the comparison of *machine learning* and deep learning models, thus establishing the best model to use for a particular time series problem. A new technique based on the spatio-temporal properties of the data has been proposed to infer information from areas for which no data are available. On the other hand, optimization problems have been formalized with up to four objectives, based on filter and wrapper methods, to perform attribute selection. It has been studied which is the best multi-objective evolutionary algorithm between *NSGA-II*, *NSGA-III*, *MOEA/D*, *SPEA2*, *IBEA*, ϵ-*NSGA-II* and ϵ-*MOEA* and the performance of predictions with one and with multiple surrogate models based on a LSTM. A total of four new surrogate-assisted evolutionary methods have been proposed, of which two use approaches without evolution control, and the other two are techniques with fixed generation-based evolution control.

All experiments have been performed mainly on two public time series datasets, one for the forecast of air quality, extracted from the air quality portal of the Region of Murcia (Spain) and the other one for the forecast of indoor temperature in a domotic house in Valencia (Spain) extracted from the *UCI Machine Learning Repository*. All the results have been validated with several statistical tests, including the *Diebold-Mariano* test, specific for time series. The proposed techniques have also been compared with other state-of-the-art methods of filter, wrapper and embedded types. *Python* has been the programming language for the implementation of the different methods presented in this work. For the evolutionary algorithms part, a library called *Platypus* has been used.

In conclusion, the proposed methods have obtained more accurate and robust forecasts with a lower computational cost than other state-of-the-art techniques for classification and regression problems with time series. In addition, the feature selection process has significantly reduced the initial number of attributes, simplifying the predictive deep learning models. The reduction of the computational cost of machine learning algorithms is a topic of great importance nowadays. This has been conveyed by international and national governmental institutions in their R&D strategies, such as the European Commission and the Government of Spain, with the aim of, on the one hand, reducing the carbon footprint to curb climate change, and on the other hand, increasing the explainability of machine learning models and, in general, the confidence in Artificial Intelligence. As future works, the application of the proposed methods to the prediction of other harmful compounds in the air is considered. Also, the use of multi-objective evolutionary algorithms to search for the optimal architecture for predictive deep learning models, or the inclusion of the proposed techniques in other types of deep learning architectures for application in other fields such as computer vision and natural language processing.

Chapter 1

Introduction

Time series are a type of data whose main differentiating characteristic is that they have a time component. They consist of a set of data taken at regular or irregular intervals and ordered chronologically. Time series have been used in a large number of domains, such as financial markets [1], internet of things [2], health care [3], stock markets [4], air quality [5], and many others. The applications of time series are very diverse, ranging from *pattern recognition* [6] to *forecasting* [7]. *Machine learning* and *deep learning* methods can be used for classification and forecasting of time series problems. However, due to the increasing amount of data being collected, one of the main problems of time series is their high dimensionality and complexity that can be present in the relationships between attributes. This is known as the *curse of dimensionality* [8]. This issue can be even worse in the case of using non-linear models, which may cause overfitting or instability in the model [9]. One way to mitigate the curse of dimensionality is through feature selection.

Feature selection (FS) [10] is a process in which relevant features are selected from a data set, thus discarding redundant and/or noisy information. This allows the reduction of the input dimensions of the data, as well as the complexity of the models created from those reduced data. It also improves the generalization capability of the models to previously unseen data, leading to more robust and accurate models. There are three main methods within FS: *filter*, *wrapper* and *embedded*. Filters separate the attribute selection from the learning algorithm so that the influence of the learning algorithm does not interact with the attribute selection algorithm. They are usually statistical measures of information. Wrappers use the predictive accuracy of a predetermined learning algorithm to determine the quality of the selected attributes. Embedded achieve model fitting and feature selection simultaneously. While filters and embedded are computationally fast methods, wrappers are very expensive. Nevertheless, wrapper methods tend to find better combinations of attributes.

Multi-objective evolutionary algorithms (MOEAs) [11] are multi-objective global search and optimization techniques inspired by the mechanisms of Darwin's natural selection and genetics. They have been successfully applied in recent years for FS [12, 13]. Multi-objective optimization problems are characterized by conflicting objective functions. Their purpose is to find the best balance or trade-off between the objective functions. This balance establishes a set of solutions called *non-dominated solutions*, which forms the

Pareto front. The Pareto front introduces the concept of *dominance* relationship between the objectives of the set of solutions. One solution is said to dominate another when one possible solution A is at least equal than another possible solution B in all objectives and solution A is better than solution B in at least one objective. MOEAs are able to successfully approximate the Pareto front due to their population-based nature. The use of wrapper methods for FS can be formulated as a multi-objective optimization problem. However, they require a high computational cost to reach a set of adequate non-dominated solutions. To avoid this, *surrogate* models can be used to approximate the fitness function of a MOEA [14]. Surrogate models, also known as meta-models, simulate the behaviour of a model and try to approximate its results as closely as possible, which makes possible a reduction in computational costs and has been successfully applied in conjunction with MOEAs [15, 16].

The main area of application of the techniques proposed in this thesis is air quality prediction. It is undeniable that the emission of certain gases into the atmosphere, such as carbon dioxide (CO_2), nitrogen dioxide (NO_2), nitrogen oxides (NO_X), or particulate matter (PM), is deteriorating air quality at a high rate. The World Health Organization estimates that in 2022 99% of the world's population will have lived in areas where the recommended levels of air quality are not attained [17]. Prolonged exposure to these noxious gases can affect the health of human beings, causing various problems such as respiratory diseases [18], the transmission of respiratory viral infections [19] and immune system problems [20], among others. According to the European Environment Agency, in 2020 more than 600,000 premature deaths occurred in European countries due to the concentration of pollutants [21]. Deteriorating air quality not only affects people but can also have a negative impact on ecosystems, resulting in vegetation loss [22] or acid rain. The creation of predictive air quality models will allow the establishment and adoption of the necessary measures to mitigate the risks posed by the concentration of gases.

In the context of the doctoral thesis, an initial work was carried out for the search, training and validation of suitable architectures for air quality time series. Thus, a study of different machine learning and deep learning techniques has been carried out. These architectures have been subsequently used in the design of MOEAs for FS. Several optimization models with different number of objective functions have been studied to determine the most relevant ones as well as the behavior of several surrogate-assisted approaches. The spatio-temporal characteristics in the use of MOEAs and the application of incremental learning for updating surrogate models have also been investigated. All these techniques have been applied mainly to two public datasets, one related to air quality and the other to the temperature inside a domotic house. In the first case, the data have been extracted from the *Autonomous Community of the Region of Murcia (Spain)*[1] and in the second from the *UCI Machine Learning Repository* [23][2]. All proposed methods have been compared with other state-of-the-art techniques. In addition, during the PhD stay at Siemens AG Österreich (Austria), a complementary analysis to the thesis studies in the area of time series classification and clustering has been carried out. For this purpose, a dataset with consumption measurements from sensors in intelligent buildings has been used.

[1] `https://sinqlair.carm.es/calidadaire/redvigilancia/redvigilancia.aspx`

[2] `https://archive.ics.uci.edu/ml/datasets/SML2010`

The remainder of the doctoral thesis is organized as follows. Section 2 describes the main objectives of this thesis. Section 3 summarizes the research methodology and details the results derived from the experiments developed within this thesis. Section 4 presents the conclusions and future work. Finally, Section 5 includes the publications that are part of the thesis by compendium.

Chapter 2

Global summary of research objectives

This section details the objectives of the PhD thesis. A general objective and six specific objectives have been proposed, which unify the partial objectives proposed in each of the works of the compendium. The general objective proposed is the following:

> Develop efficient and effective feature selection techniques for deep learning through multi-objective evolutionary algorithms and application of the created methods for time series forecasting in different areas of interest.

This general objective has been broken down into the following specific objectives:

SO1: Develop a comprehensive methodology and implement a multi-criteria decision-making process for the comparison and evaluation of predictive models for time series forecasting.

SO2: Study, design and develop a multi-objective evolutionary approach based on spatio-temporal characteristics within the Autonomous Region of Murcia.

SO3: Define multi-objective optimization problems for feature selection, with objectives of different nature, both filter and wrapper.

SO4: Solve the proposed optimization problems by identifying the best state-of-the-art multi-objective evolutionary algorithms and developing surrogate-assisted approaches to reduce the computational cost of the algorithms.

SO5: Identify metrics to quantify the variability between surrogate-assisted approaches and facilitate the establishment of qualitative analysis.

SO6: Evaluate, validate and compare the developed feature selection methods with time series data for air quality forecasting in the context of the Autonomous Region of Murcia, as well as in other geographic locations and in other time series forecasting problems for the sake of generalization verification.

The following chapter shows the methodology carried out to meet the proposed objectives, as well as the main results obtained.

Chapter 3

Summary of research methodology and results

The methodology and results of this thesis can be found in the articles of the above-mentioned compendium of publications. This section summarizes the followed methodology and the most relevant results related to the objectives of the thesis. Additionally, the datasets used to validate the experiments are briefly described.

3.1 Datasets

During the course of the thesis, multiple datasets have been used to evaluate the proposed methods. They mainly comprise two types, air quality and indoor temperature.

The first dataset, with which the first article was developed, contains air quality data from a monitoring station located in the city of Wroclaw, Poland. Data were collected between 2015 and 2017. Initially, the data had 26304 instances and 9 attributes and the output variables are NO_2 and NO_X.

For the second paper, four datasets belonging to four monitoring stations within the Region of Murcia, Spain, have been used. The stations are located in Alcantarilla, La Aljorra, Lorca and Valle de Escombreras and take daily air quality information between 2017 and 2020. In total, there are 1461 instances and 19 attributes, including the latitude and longitude of the monitoring stations. The output variable is NO_2.

The rest of the articles have used two datasets, one for air quality at La Aljorra described above and the other for indoor temperature. The indoor temperature dataset was initially used in [24] and contains data taken every 15 minutes during March and May 2012 in a domotic house located in Valencia, Spain. In total, it has 4137 instances and 24 attributes. The output is the temperature inside the dining room.

The time series dataset used in the PhD stay in Vienna contains hourly data on energy, hot water, cold water and heating consumption in different smart buildings. This dataset is private and belongs to Siemens AG Österreich, so a detailed description of the data can not be provided, but the results obtained from the data can be presented.

3.2 Methodology for the identification of deep learning architectures and comparison of predictive models

For the identification of deep learning architectures and to establish a fair comparison between different predictive models, regardless of whether they are machine learning or deep learning, the methodology described in Figure 3.1 has been proposed in [25]. This methodology is generic, as it can be adapted to any type of regression problem and does not rely on any inherent learning algorithm. Moreover, by being able to compare various *window sizes* (WS), it is possible to find the most appropriate dataset for reliable and robust predictions. The phases of the methodology are as follows: sliding window transformation with different window sizes, hyper-parameter tuning with 3-fold cross-validation and 1 repetition, statistical tests performed with 10-fold cross-validation with 3 repetitions, multi-criteria decision making for *mean absolute error* (MAE), *root mean square error* (RMSE) and *correlation coefficient* (CC) and step ahead predictions.

For the time series problem on which this methodology has been validated, the following learning algorithms have been used: *1 dimension convolutional neural network* (1D-CNN) [26], *gated recurrent unit* (GRU) [27], *long short-term memory* (LSTM) [28], *random forest* (RF) [29], *lasso* [30], and *support vector machine* (SVM) [31] with radial kernel. The multi-criteria decision-making process takes into consideration the exactness and robustness of the MAE, RMSE and CC of the models, and gathers these criteria into a single weighted metric called *goodness*. The model with the lowest goodness is the one that is finally selected.

This methodology has been applied four times, as the original air quality dataset has been modified according to the output attribute to be predicted and the concentration of O_3 has been removed to check its impact on the predictions. Figure 3.2 represents the best architecture (LSTM-WS24) found for NO_2 prediction. Table 3.1 shows the train and test errors of the final architecture with 24-steps ahead predictions with the best models. Finally, Figure 3.3 shows the original and the predicted time series for some steps ahead.

It has been observed that for predicting NO_2 and NO_X concentrations the models based on LSTM and GRU have obtained better results compared to the rest. This result was expected, given that both LSTM and GRU are *recurrent neural networks* (RNN), i.e., they consider previous information to predict future information. This factor is the main differentiator of these learning algorithms in comparison with the rest. The 24-hour predictions for NO_2 have been very favorable, so this methodology is likely to be included in a decision system to alert the population of the increase of polluting gases in the environment.

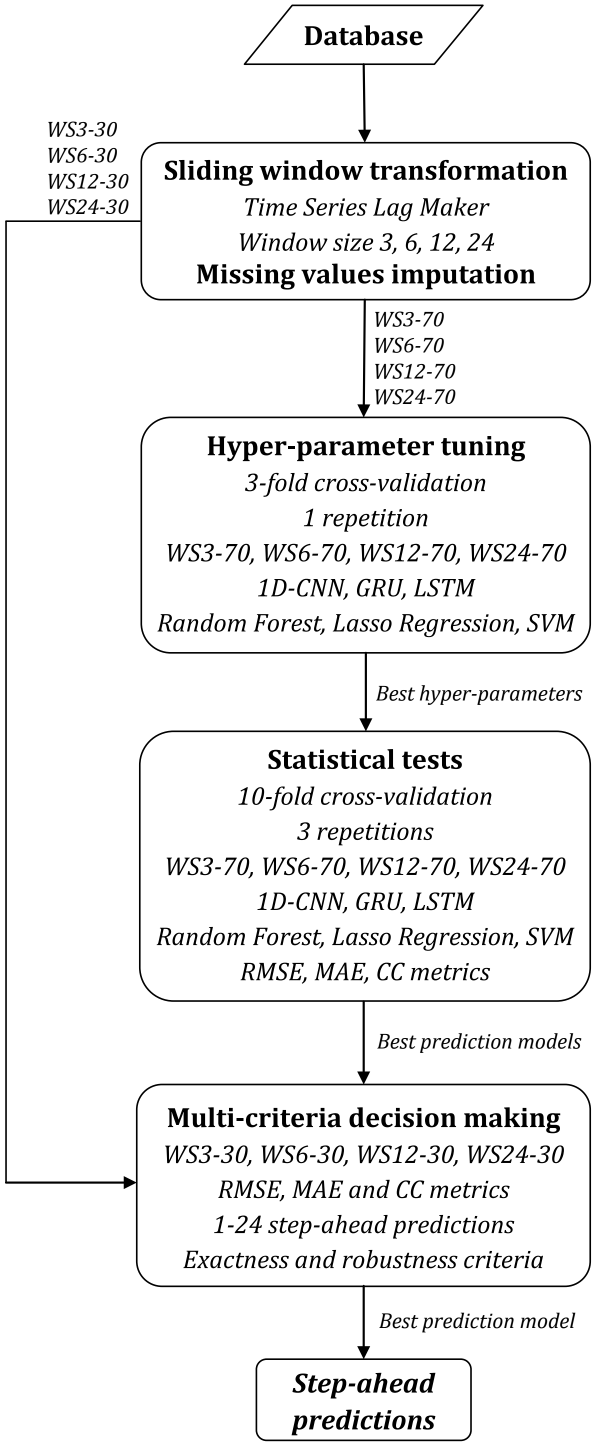


Figure 3.1: Proposed methodology for the prediction of air quality time series.

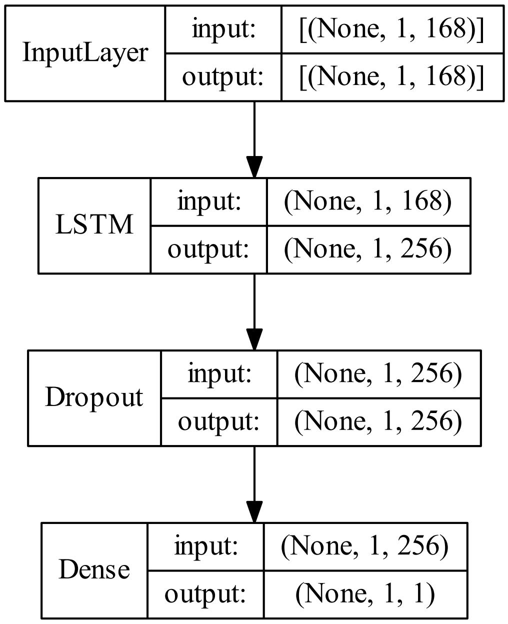


Figure 3.2: Architecture of the LSTM-WS24 deep learning model for NO_2 prediction.

		Steps-ahead											
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Training	MAE	6.730	8.344	9.360	9.972	10.332	10.568	10.723	10.840	10.927	11.008	11.092	11.155
	RMSE	9.341	11.694	13.064	13.843	14.314	14.608	14.806	14.955	15.078	15.196	15.307	15.394
	CC	0.922	0.877	0.848	0.832	0.823	0.818	0.815	0.813	0.811	0.810	0.808	0.807
Test	MAE	6.267	7.666	8.445	8.902	9.167	9.320	9.409	9.470	9.510	9.536	9.569	9.598
	RMSE	8.523	10.464	11.528	12.114	12.443	12.637	12.743	12.810	12.855	12.896	12.944	12.992
	CC	0.923	0.883	0.857	0.843	0.835	0.830	0.828	0.826	0.825	0.824	0.823	0.822

		Steps-ahead											
		13	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**
Training	MAE	11.211	11.264	11.323	11.388	11.453	11.504	11.544	11.575	11.606	11.642	11.699	11.801
	RMSE	15.469	15.535	15.608	15.690	15.770	15.838	15.895	15.939	15.979	16.026	16.107	16.259
	CC	0.806	0.805	0.805	0.804	0.803	0.802	0.801	0.801	0.800	0.800	0.799	0.796
Test	MAE	9.628	9.653	9.674	9.694	9.714	9.735	9.756	9.772	9.789	9.800	9.824	9.860
	RMSE	13.036	13.075	13.110	13.146	13.181	13.209	13.232	13.251	13.269	13.288	13.321	13.383
	CC	0.821	0.820	0.819	0.818	0.818	0.817	0.817	0.816	0.816	0.816	0.815	0.814

Table 3.1: Performance on training and test data of the LSTM-WS24 model for 24-steps-ahead predictions in the NO_2 problem.

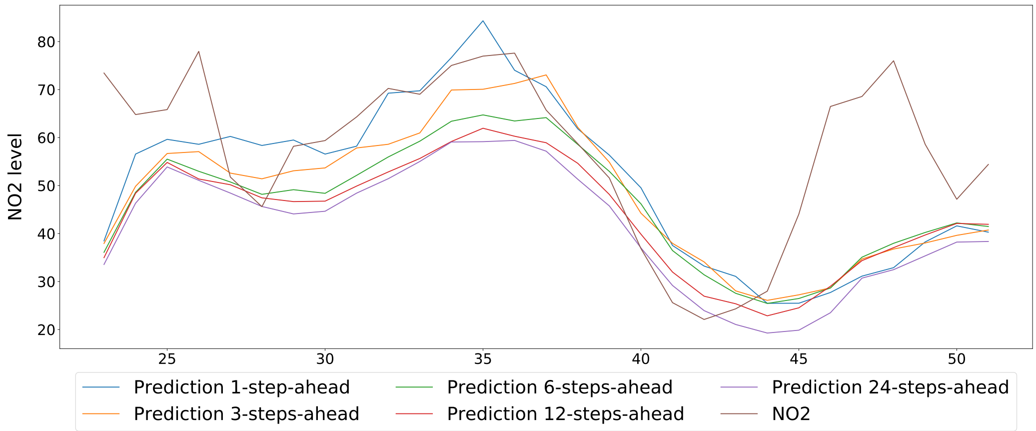


Figure 3.3: NO_2 predicted time series in test with the best architecture (LSTM-WS24).

3.3 Multi-objective optimization based spatio-temporal approach

Continuing with the study of air quality prediction, a multi-objective optimization-based spatio-temporal approach has been developed in [32]. For this purpose, the problem has first been formalized as a multi-objective optimization problem. Thus, each objective is the RMSE of a *linear regression* (LR) based predictive model of a monitoring station. These objectives try to be minimized to achieve the best predictions at each location. In this particular case, it is a 3-objective optimization problem. *Non-dominated sorting genetic algorithm II* (NSGA-II) [33], *multi-objective evolutionary algorithm based on decomposition* (MOEA/D) [34] and *strength Pareto evolutionary algorithm* (SPEA2) [35] were evaluated as MOEAs, and NSGA-II was finally selected as it presents better optimality and diversity. The Pareto fronts resulting from the genetic algorithm (as in Figure 3.5) are the input to build an ensemble learning model. The ensemble learning approach is based on *stacking*. In this work, the following learning algorithms have been used to train the ensemble model: RF, LR, SVM, *quasi-recurrent neural networks* (QRNN) [36], *multilayer perceptron* (MLP) [37], *k-nearest neighbors* (kNN) [38] and ZeroR. With the trained ensemble, the multi-step forecasting process is performed to make the predictions of the monitoring station data not seen by the genetic algorithm. In this way, a method capable of making approximations in the predictions based on nearby geographical points is obtained. Figure 3.4 shows graphically the flow that this process follows.

The proposed method has achieved better results in test than other similar approaches in the literature, such as the case of interpolation based on an *inverse distance weighting* (IDW) [39] function of the model predictions, as shown in Table 3.2. LR has been the learning algorithm with the best goodness in test, followed by QRNN. For the ensemble evaluated in the test set, our method presents a better generalization error, since for previously unseen data (those of the La Aljorra monitoring station) it manages to make better predictions. Table 3.3 shows the results of the 7-steps ahead evaluation of the LR model in test of the proposed method (MOEA + Ensemble) and the IDW-based interpolation method (Interpolation). Figure 3.6 depicts the predicted time serie for 1-

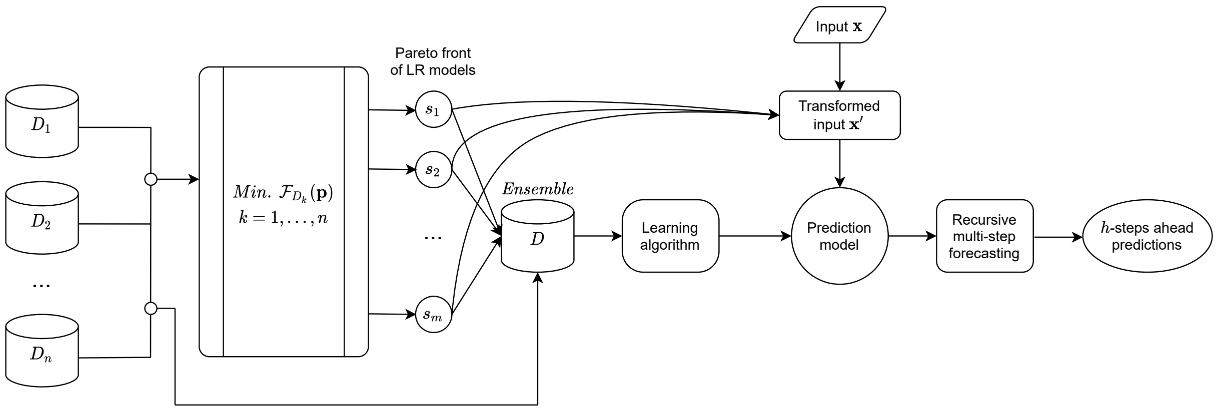


Figure 3.4: Multi-objective evolutionary optimization based spatio-temporal with ensemble learning approach.

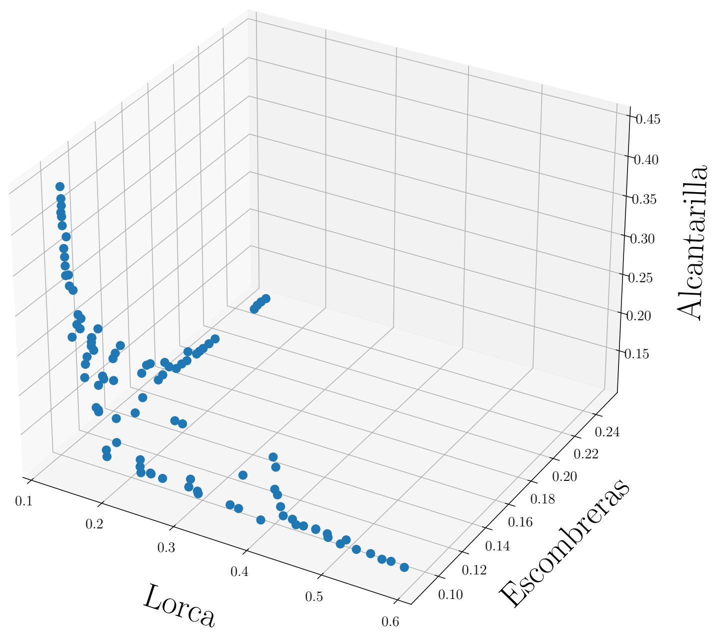


Figure 3.5: Pareto front in 3D with NSGA-II.

step ahead in test of the LR model. This new technique also shows a lower propensity to overfitting. It has been statistically demonstrated that the proposed ensemble approach is able to obtain significant differences compared to the union of datasets.

Models	**Goodness RF**	**Goodness LR**	**Goodness SVM**	**Goodness QRNN**
Training MOEA + Ensemble	0.130603	0.218094	0.206317	0.173174
Training Interpolation	0.081890	0.218155	0.136968	0.151813
La Aljorra Test MOEA + Ensemble	0.269400	0.250435	0.282549	0.267744
La Aljorra Test Interpolation	0.339744	0.326777	0.329596	0.302130

Table 3.2: Goodness of the predictions models with RF, LR, SVM and QRNN.

Method	**Metric**	**1-step ahead**	**2-steps ahead**	**3-steps ahead**	**4-steps ahead**	**5-steps ahead**	**6-steps ahead**	**7-steps ahead**
MOEA + Ensemble	RMSE	0.1118	0.1207	0.1229	0.1233	0.1207	0.1261	0.1408
	MAE	0.0890	0.0960	0.0980	0.0984	0.0961	0.1014	0.1154
	CC	0.5410	0.4863	0.4745	0.4724	0.4860	0.4581	0.3830
Interpolation	RMSE	0.1175	0.1207	0.1212	0.1296	0.1322	0.1346	0.1352
	MAE	0.0919	0.0936	0.0939	0.1014	0.1027	0.1035	0.1035
	CC	0.3199	0.2673	0.2393	0.2420	0.2574	0.2057	0.1874

Table 3.3: Results of the evaluation of models on La Aljorra test set with LR.

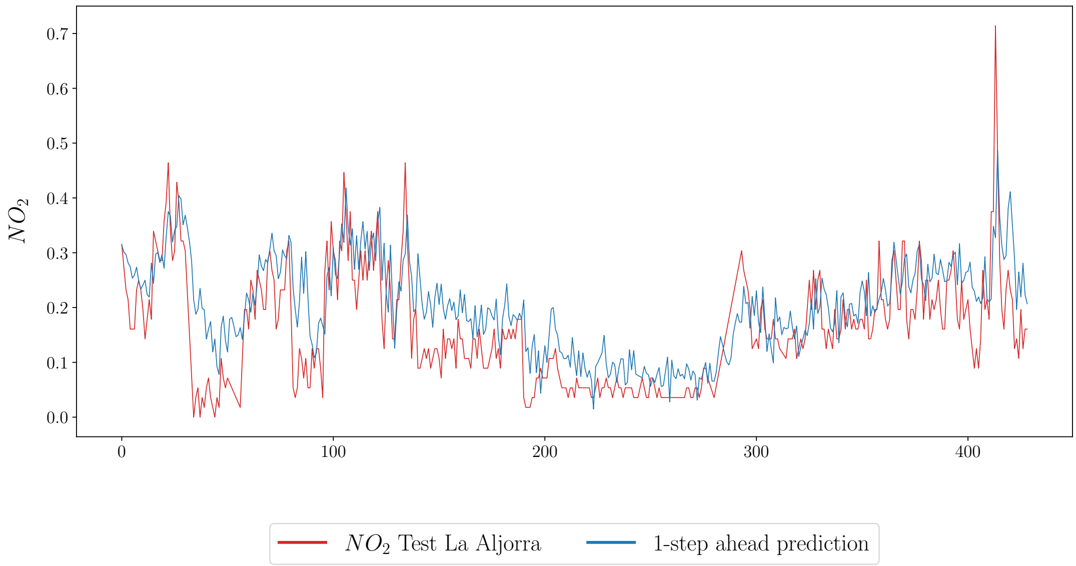


Figure 3.6: Original and predicted NO_2 time series for 1-step ahead built with the multi-objective optimization based spatio-temporal approach for LR.

3.4 Multi-objective evolutionary algorithms based on surrogate models

After analyzing the most appropriate state-of-the-art learning algorithms for time series and the performance of MOEAs, several techniques based on surrogate methods have been developed. These techniques are described in the following sections.

3.4.1 Surrogate-assisted and filter-based multi-objective evolutionary FS for deep learning

A multi-objective FS method based on surrogate-assisted models for deep learning has been presented in [40]. Up to four objectives have been defined to formalize the FS problem with wrapper, filters and hybrids as a multi-objective optimization problem. *Correlation* [41] and *reliefF* [42] algorithms have been adapted for attribute subset evaluation and used to define the objectives related to the filter methods. Figure 3.7 shows a flow diagram of the surrogate-assisted multi-objective evolutionary algorithm.

A generic FS method applicable to any regression problem has been developed and tested on air quality and indoor temperature time series. Thanks to a surrogate model, this method allows the use of deep learning models as the learning algorithm of a wrapper method, but without the disadvantage of the high computational cost that this would have. The surrogate-assisted model also helps to reduce possible overfitting and achieve a better generalization capacity. The idea behind the proposed approach is the following: the surrogate model is evaluated with a test dataset consisting of the original test dataset where the values of the unselected attributes are set to a constant value $\alpha = 0$ in all samples. When an attribute is redundant or irrelevant, the evaluation of the surrogate model, which has been trained on all the attributes, is little impaired by this change in the test data set. However, when an attribute is relevant, the evaluation of the surrogate model will be greatly altered since the attribute has been highly influential in the training of the surrogate model. Several state-of-the-art MOEAs have been compared, including NSGA-II, *non-dominated sorting genetic algorithm III* (NSGA-III) [43], MOEA/D, SPEA2, *indicator-based evolutionary algorithm* (IBEA) [44], ϵ-*MOEA* [45] and ϵ-*NSGA-II* [46]. Figure 3.8 shows the Pareto front for NSGA-II.

The new multi-criteria summary performance metric, $\mathcal{H}$, allows adjusting the importance of the metrics that form it, being able to give more weight to the one that is more important for the treated problem but without losing the additional information provided by the rest of the metrics. It makes it possible to compare different models and their predictions at h-steps ahead. The metric is independent of the prediction horizon so that the most appropriate one for each problem can be used. In addition, it allows obtaining more robust predictive models, since it considers several metrics and summarizes them. The final results of the best predictive model found for 7-steps ahead in train and test are shown in Table 3.4 as well as the time series of the predicted NO_2 values for 1-step ahead in Figure 3.9.

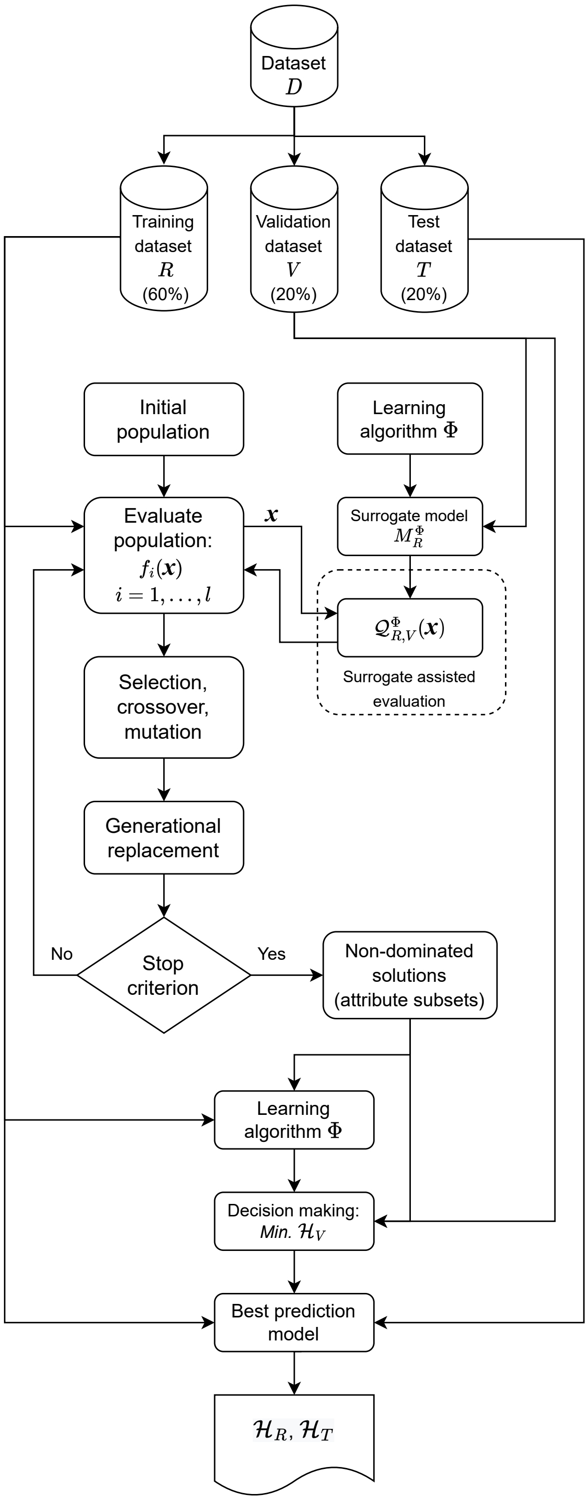


Figure 3.7: Surrogate-assisted multi-objective evolutionary algorithm for FS.

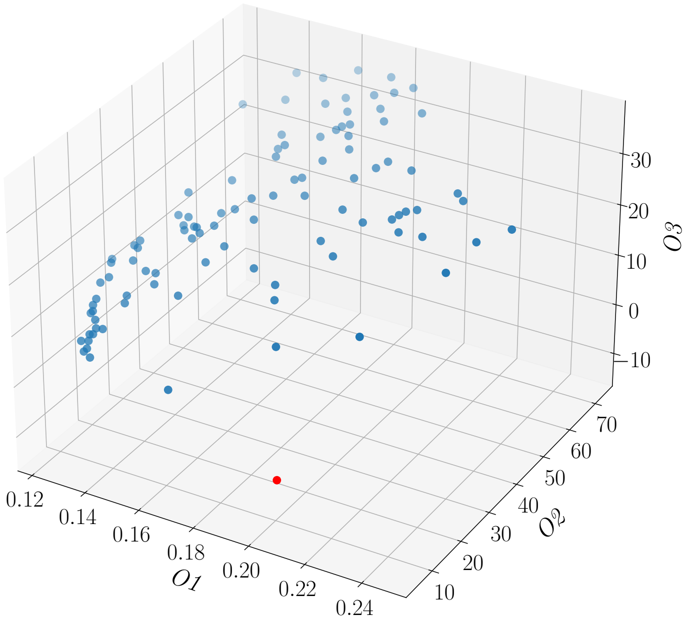


Figure 3.8: Pareto front in 3D for the best prediction model (*O1O2O3*) and the best MOEA (NSGA-II) for air quality. Red point represents the model with best average RMSE of 7-steps ahead predictions.

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

Table 3.4: Results of the best prediction model (obtained with *O1O2O3-NSGA-II* method) for air quality evaluated on the training set R and test set T datasets.

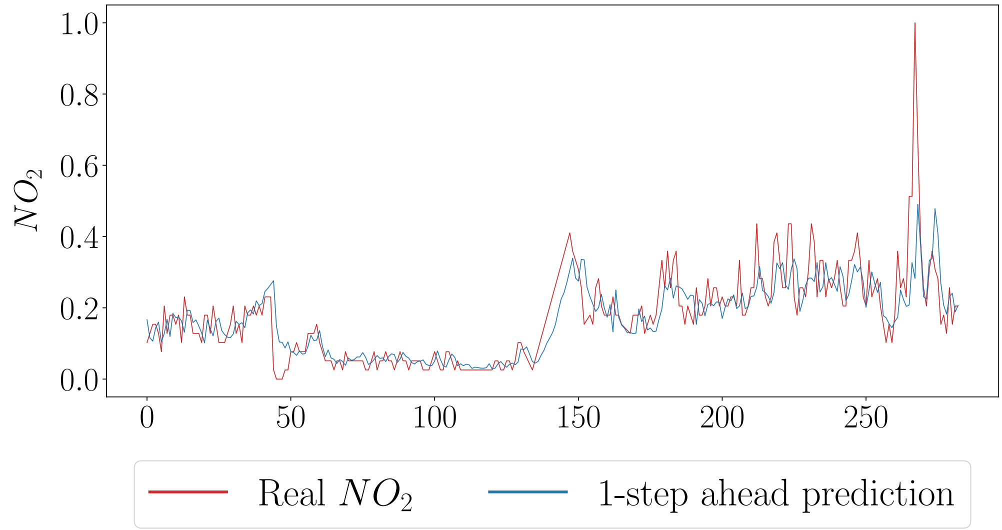


Figure 3.9: Times series of 1-step ahead predictions for NO_2 evaluated on test T of the prediction model obtained with *O1O2O3-NSGA-II*.

In both problems studied, a reduction of the attributes of more than 80% for the initial set has been achieved, making the new reduced datasets more interpretable. Furthermore, it has been determined that all the defined objectives are relevant in the multi-objective search for the best subset of features. In the comparison with other FS methods in the literature (Table 3.5), the proposed method outperforms most of them in terms of $\mathcal{H}_R$ ($\mathcal{H}$ in the train set R) and all of them in terms of $\mathcal{H}_T$ ($\mathcal{H}$ in the test set T). The literature methods compared have been: hybrid filter-wrapper FS method based on correlation and LSTM with deterministic search (*M1*), hybrid filter-wrapper FS method based on reliefF and LSTM with deterministic search (*M2*), wrapper multi-objective evolutionary FS method based on linear regression (*M3*), wrapper multi-objective evolutionary FS method based on random forest (*M4*), *CancelOut* [47] (*M5*) and random forest (*M6*).

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 3.5: Comparison of feature selection methods for the air quality problem, sorted from best to worse evaluation of $\mathcal{H}_T$.

3.4.2 Multi-surrogate assisted multi-objective evolutionary algorithm for FS with deep learning

In [40] it has been proven that the use of surrogate models is effective for FS processes applied to regression problems. Therefore, the next step was to test the behavior of multiple surrogate-assisted models [48], in order to achieve better generalizability of the predictive models. For this purpose, a scheme similar to Figure 3.7 has been followed but incorporating multiple surrogate models based on a deep learning algorithm and a new metric to decide the best model. This new algorithm is shown in Figure 3.10. In addition, the effectiveness of the proposed method has been validated in both regression and classification problems.

A new variability metric named $\mathcal{V}_X$ has been defined to qualitatively analyze the FS results. Thus, it is compared whether the order of the results of the multi-surrogate method is similar to those of a wrapper method. The results show that in both cases the worst subset of attributes coincides. Moreover, for both regression and classification, the variability does not exceed 36% in any case, so it can be determined that the proposed approach is reliable.

For the air quality problem, the multi-surrogate method obtains better results in $\mathcal{H}_T$ than the conventional wrapper method at the same run time, both for regression and classification. This evidences the fact that the multi-criteria metric $\mathcal{H}$ can be extended as a classification metric. Thus, comparisons between different models can be successfully established. At the statistical level, the proposed method obtains significant differences in comparison with the wrapper method. Moreover, the predictions are more stable and, therefore, robust. Tables 3.6 and 3.7 show the results of the 7-steps ahead best model predictions for the train and test sets for the regression and classification problem, respectively. Figures 3.11 and 3.12 represent the 7-steps ahead predictions for NO_2 for the best models in regression and classification. In terms of diversity, as shown in Figures 3.13 and 3.14 the Pareto fronts belonging to the multi-surrogates are more diverse, obtaining a greater variety of non-dominated solutions. To demonstrate the generalizability of the proposed method, it has also been applied to an indoor temperature problem, with satisfactory results. Finally, it has been compared with two other state-of-the-art FS methods, CancelOut and lasso, outperforming their results.

Evaluation dataset	**Performance metric**	**1-step ahead**	**2-steps ahead**	**3-steps ahead**	**4-steps ahead**	**5-steps ahead**	**6-steps ahead**	**7-steps ahead**
	RMSE	0.0780	0.0790	0.0801	0.0805	0.0807	0.0810	0.0820
R	MAE	0.0549	0.0558	0.0569	0.0573	0.0576	0.0580	0.0586
	CC	0.8469	0.8395	0.8345	0.8313	0.8293	0.8269	0.8235
	RMSE	0.0633	0.0715	0.0746	0.0758	0.0764	0.077	0.0795
T	MAE	0.0435	0.0493	0.0528	0.0545	0.0552	0.0554	0.0578
	CC	0.6967	0.6052	0.5687	0.5533	0.5474	0.5404	0.5068

Table 3.6: RMSE, MAE and CC of the multi-step ahead forecasting for the multi-surrogate assisted multi-objective evolutionary algorithm with LSTM (air quality regression problem), evaluated on the training set R and the test set T.

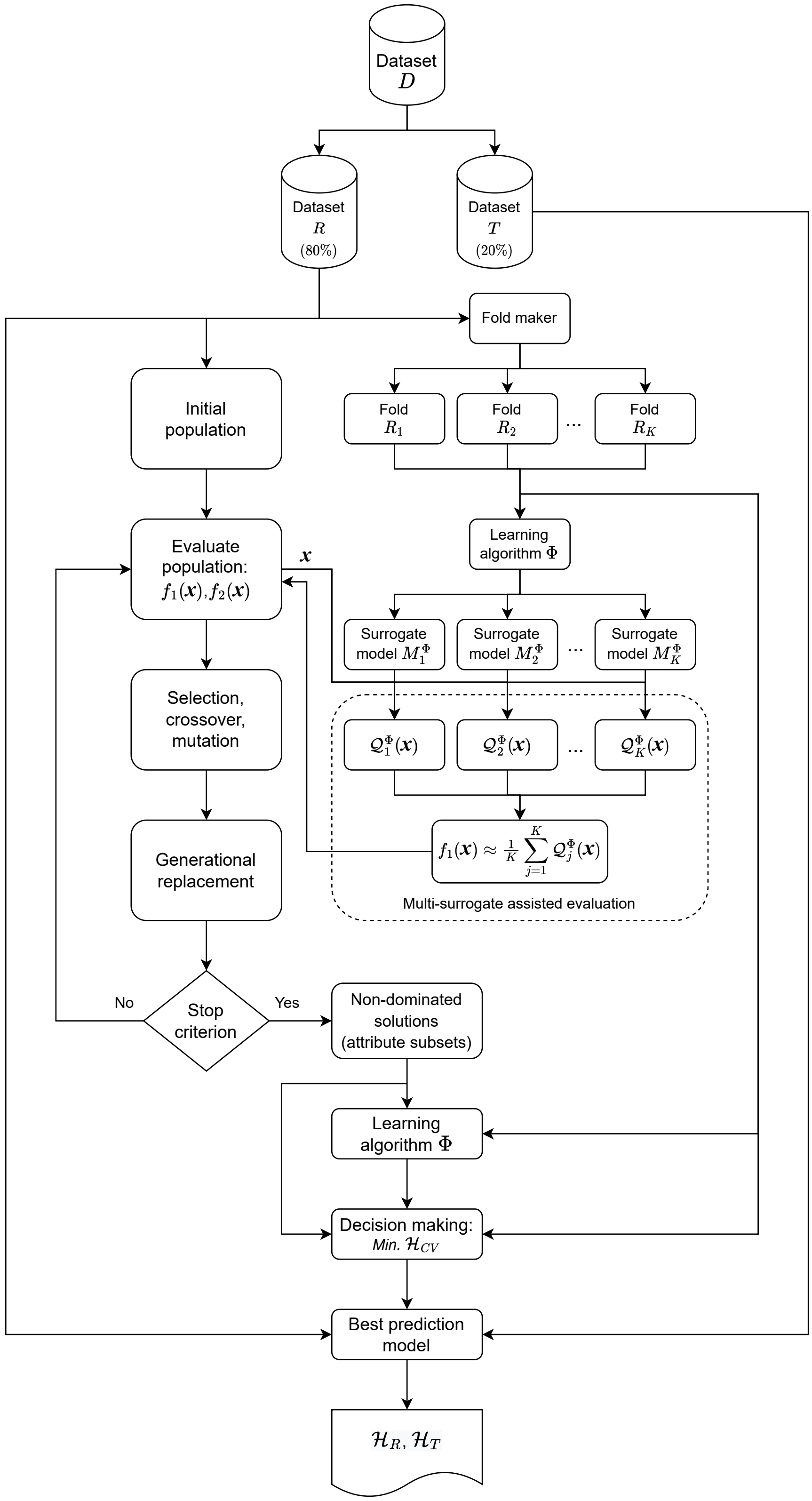


Figure 3.10: Multi-surrogate-assisted multi-objective evolutionary algorithm for FS.

Evaluation dataset	**Performance metric**	**1-step ahead**	**2-steps ahead**	**3-steps ahead**	**4-steps ahead**	**5-steps ahead**	**6-steps ahead**	**7-steps ahead**
R	BA	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	AUC	1.0	1.0	1.0	1.0	1.0	1.0	1.0
T	BA	0.3270	0.3209	0.3153	0.3158	0.3163	0.3195	0.3212
	AUC	0.8063	0.7911	0.7770	0.7779	0.7787	0.7862	0.7882

Table 3.7: BA and AUC of the multi-step ahead forecasting for the multi-surrogate assisted multi-objective evolutionary algorithm with RF (air quality classification problem), evaluated on the training set R and the test set T.

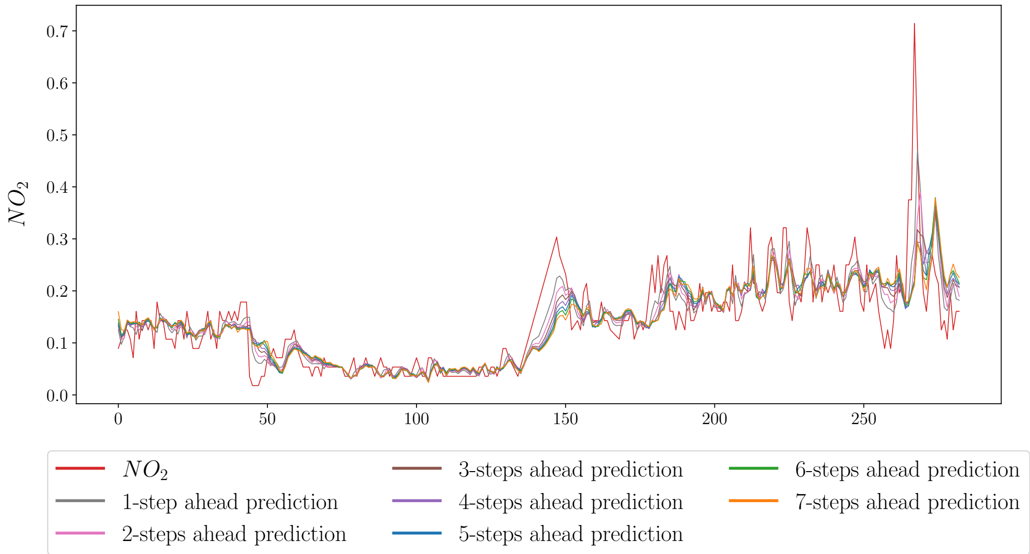


Figure 3.11: 7-steps ahead forecasting for NO_2 of the multi-surrogate assisted MOEA with LSTM for regression evaluated on test.

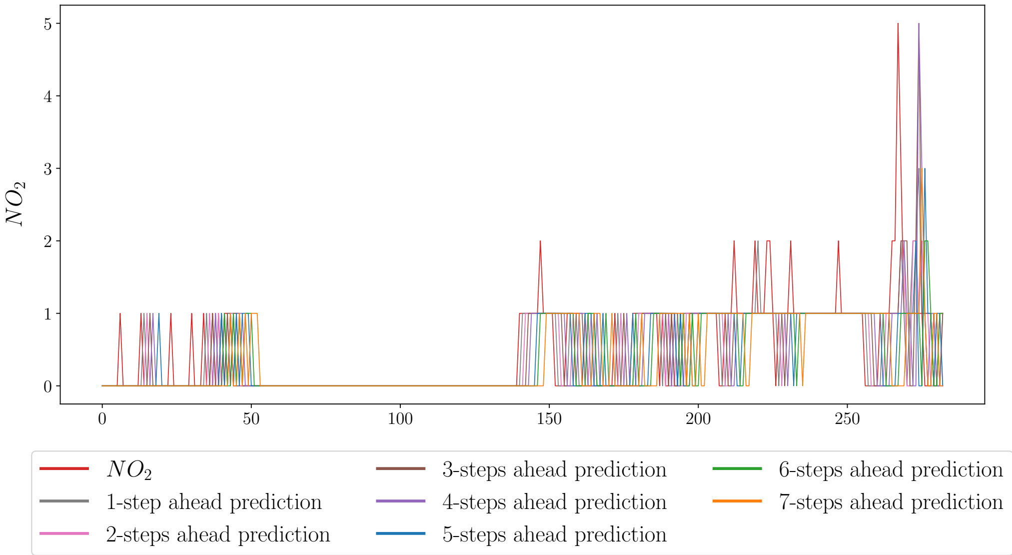


Figure 3.12: 7-steps ahead forecasting for NO_2 of the multi-surrogate assisted MOEA with RF for classification evaluated on test.

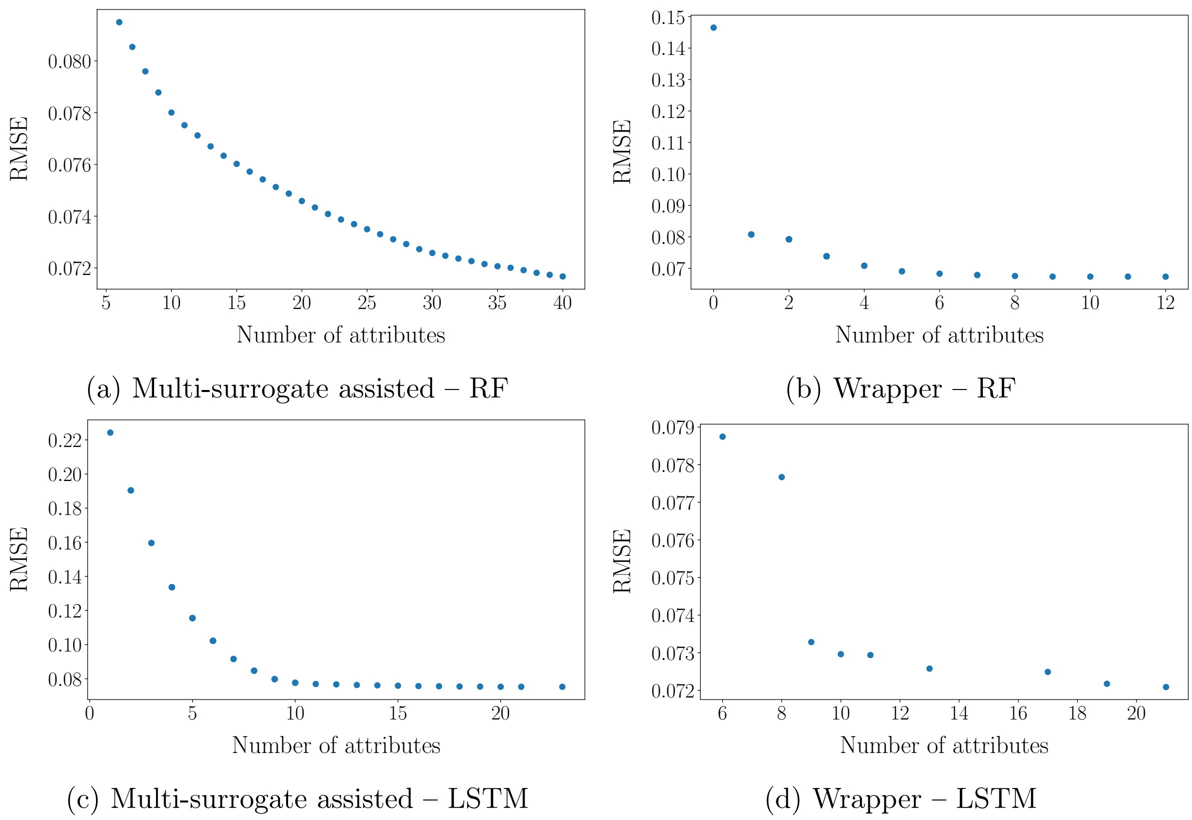
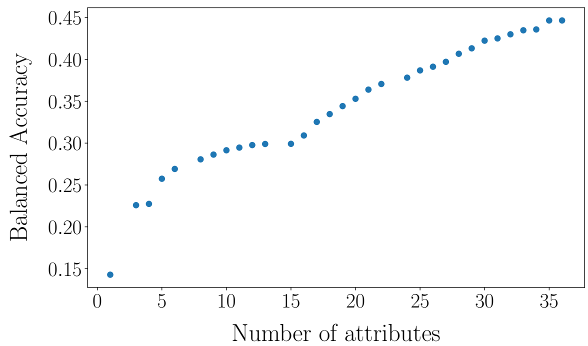
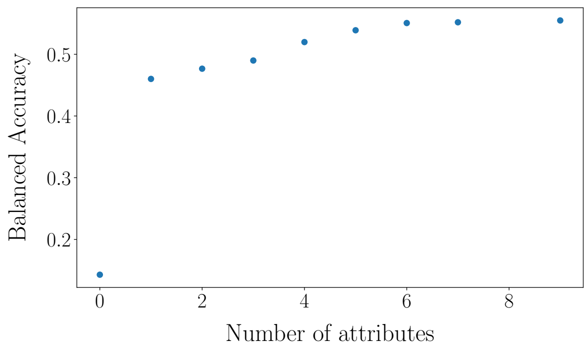


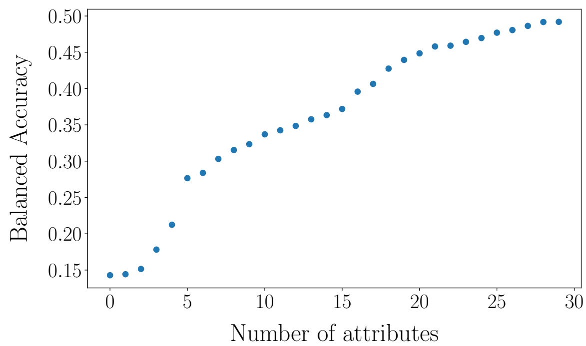
Figure 3.13: Pareto fronts obtained with NSGA-II with the multi-surrogate assisted and wrapper approaches for the air quality regression problem.



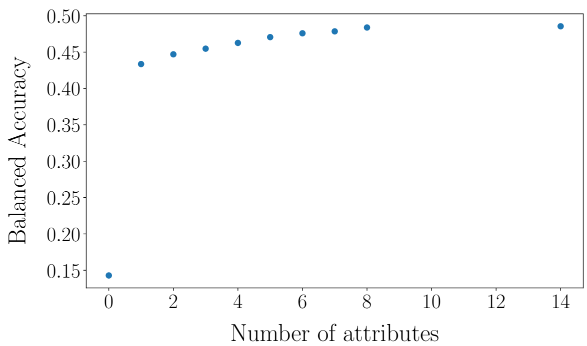
(a) Multi-surrogate assisted – RF



(b) Wrapper – RF



(c) Multi-surrogate assisted – SVM



(d) Wrapper – SVM

Figure 3.14: Pareto fronts obtained with NSGA-II with the multi-surrogate assisted and wrapper approaches for the air quality classification problem.

3.4.3 Surrogate-assisted multi-objective evolutionary algorithm of generation-based fixed evolution control for FS with deep learning

In the previous sections, it has been shown that the use of one or more surrogate-assisted models in MOEAs are able to perform the FS process successfully. However, the surrogate model always maintains the same information as at the beginning of the method. This, in a way, is wasteful, since the underlying information obtained in the genetic algorithm is not considered. For this reason, two approaches have been proposed for updating the surrogate model, one based on *incremental learning* [49] (Algorithm 1) and the other based on updating the database and building a new surrogate model (Algorithm 2). Figure 3.15 shows the scheme followed by the surrogate-assisted multi-objective evolutionary algorithm with incremental learning for FS. In this case, the surrogate model is obtained by building offline a learning meta-model from a set of samples of attribute selections and their actual evaluations are obtained by training a neural network (LSTM) for each sample.

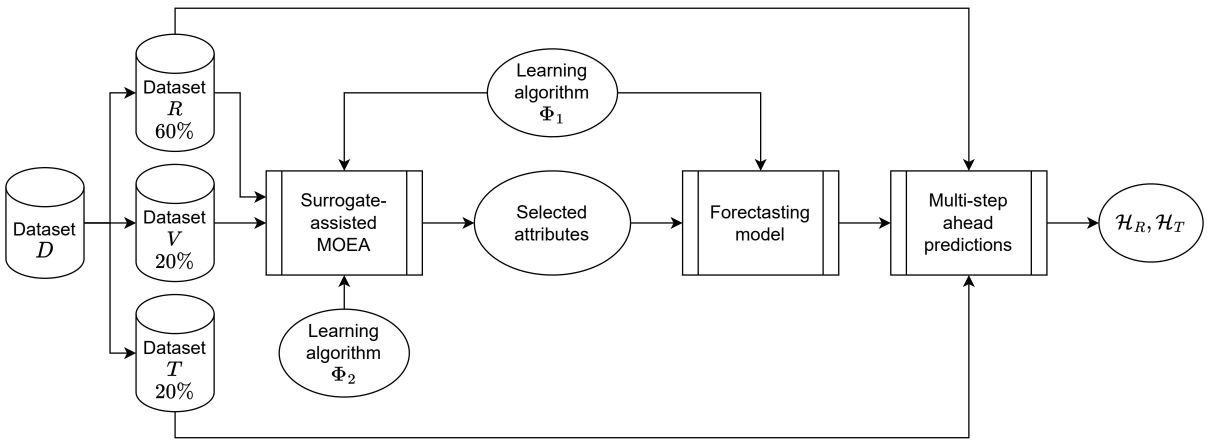


Figure 3.15: Surrogate-assisted multi-objective evolutionary algorithm with incremental learning for FS.

The method based on updating the surrogate model has achieved better results than the previously proposed method without updating the surrogate model, both for the air quality (Table 3.8) and for the indoor temperature problems. Figure 3.16 shows the NO_2 predicted time series for the proposed method evaluated on test. As shown in Table 3.9, compared to other state-of-the-art methods it also manages to achieve better results in terms of $\mathcal{H}$. The statistical tests have been performed on the h-steps ahead predictions and not on the error metrics, as was previously done. For this purpose, the *Diebold Mariano test* [50] has been applied. Thus, win-loss rankings are established to determine the statistical significance of the proposed method.

Although there is no reduction in the number of attributes for the method without updating compared to the $O1O2$ approach of [40], there is still a decrease of more than 80% in the number of attributes in contrast with the original number of features, thus,

Algorithm 1: Incremental learning based surrogate-assisted multi-objective evolutionary algorithm for feature selection

```
Data: R;                                            // Training dataset
Data: V;                                            // Validation dataset
Data: Φ_1;            // Learning algorithm of the optimization problem
Data: Φ_2;                 // Learning algorithm for surrogate model
Data: G > 1;                                   // Number of generations
Data: N > 1;               // Number of individuals in the population
Data: E;                            // Surrogate model update frequency
Data: epochs;              // Number of epochs for incremental learning
Result: FS;                                        // Feature selection
D_{R,V}^{Φ_1} ← BuiltAuxiliaryDataset(R, V, Φ_1);
M_{R,V}^{Φ_1Φ_2} ← BuiltSurrogateModel(D_{R,V}^{Φ_1}, Φ_2);
P ← InitializePopulation(N);
Approximate(P, M_{R,V}^{Φ_1Φ_2});            // Evaluation with surrogate model
g ← 1;
while g < G do
    Q ← ∅;
    i ← 1;
    while i < N/2 do
        parent1 ← Selection(P);
        parent2 ← Selection(P);
        (child1, child2) ← Crossover(parent1, parent2);
        offspring1 ← Mutation(child1);
        offspring2 ← Mutation(child2);
        Q ← Q ∪ {offspring1, offspring2};
        i ← i + 1;
    end
    Approximate(Q, M_{R,V}^{Φ_1Φ_2});        // Evaluation with surrogate model
    P ← PopulationUpdate(P, Q);
    if (g mod E) = 0 then
        P′ ← P;
        Evaluate(P′, H_V);              // Evaluation with real fitness function
        ND ← NonDominated(P′);
        NR ← NonRepeated(ND, D_{R,V}^{Φ_1});
        UpdateAuxiliaryDataSet(D_{R,V}^{Φ_1}, NR);
        UpdateSurrogateModel(M_{R,V}^{Φ_1Φ_2}, NR, epochs);
    end
    g ← g + 1;
end
FS ← DecisionMaking(ND);
return FS
```

Algorithm 2: Non-incremental surrogate-assisted multi-objective evolutionary algorithm for feature selection

Data: R; // Training dataset
Data: V; // Validation dataset
Data: Φ_1; // Learning algorithm of the optimization problem
Data: Φ_2; // Learning algorithm for surrogate model
Data: $G > 1$; // Number of generations
Data: $N > 1$; // Number of individuals in the population
Data: E; // Surrogate model update frequency
Result: FS; // Feature selection

$D^{\Phi_1}_{R,V} \leftarrow$ BuiltAuxiliaryDataset(R, V, Φ_1);
$M^{\Phi_1\Phi_2}_{R,V} \leftarrow$ BuiltSurrogateModel($D^{\Phi_1}_{R,V}$, Φ_2);
$P \leftarrow$ InitializePopulation(N);
Approximate(P, $M^{\Phi_1\Phi_2}_{R,V}$); // Evaluation with surrogate model
$g \leftarrow 1$;
while $g < G$ **do**
  $Q \leftarrow \emptyset$;
  $i \leftarrow 1$;
  while $i < N/2$ **do**
    parent1 $\leftarrow$ Selection(P);
    parent2 $\leftarrow$ Selection(P);
    (*child1*, *child2*) $\leftarrow$ Crossover(*parent1*, *parent2*);
    offspring1 $\leftarrow$ Mutation(*child1*);
    offspring2 $\leftarrow$ Mutation(*child2*);
    $Q \leftarrow Q \cup \{$*offspring1*, *offspring2*$\}$;
    $i \leftarrow i + 1$;
  end
  Approximate(Q, $M^{\Phi_1\Phi_2}_{R,V}$); // Evaluation with surrogate model
  $P \leftarrow$ PopulationUpdate(P, Q);
  if $(g \mod E) = 0$ **then**
    $P' \leftarrow P$;
    Evaluate(P', $\mathcal{H}_V$); // Evaluation with real fitness function
    $ND \leftarrow$ NonDominated(P');
    $NR \leftarrow$ NonRepeated(ND, $D^{\Phi_1}_{R,V}$);
    UpdateAuxiliaryDataSet($D^{\Phi_1}_{R,V}$, NR);
    $M^{\Phi_1\Phi_2}_{R,V} \leftarrow$ BuiltSurrogateModel($D^{\Phi_1}_{R,V}$, Φ_2);
  end
  $g \leftarrow g + 1$;
end
$FS \leftarrow$ DecisionMaking(ND);
return FS

Evaluation dataset	Performance metric	1-step ahead	2-steps ahead	3-steps ahead	4-steps ahead	5-steps ahead	6-steps ahead	7-steps ahead
R	RMSE	0.0834	0.0820	0.0818	0.0819	0.0818	0.0818	0.0815
	MAE	0.0572	0.0573	0.0575	0.0576	0.0574	0.0574	0.0573
	CC	0.8661	0.8681	0.8683	0.8668	0.8664	0.8655	0.8666
T	RMSE	0.0943	0.0975	0.0979	0.0981	0.0983	0.0984	0.0985
	MAE	0.0727	0.0760	0.0766	0.0769	0.0770	0.0771	0.0771
	CC	0.7535	0.7462	0.7469	0.7466	0.7465	0.7466	0.7467

Table 3.8: Results of the best forecast model with RF for the air quality forecast problem, evaluated on the training set R and the test set T.

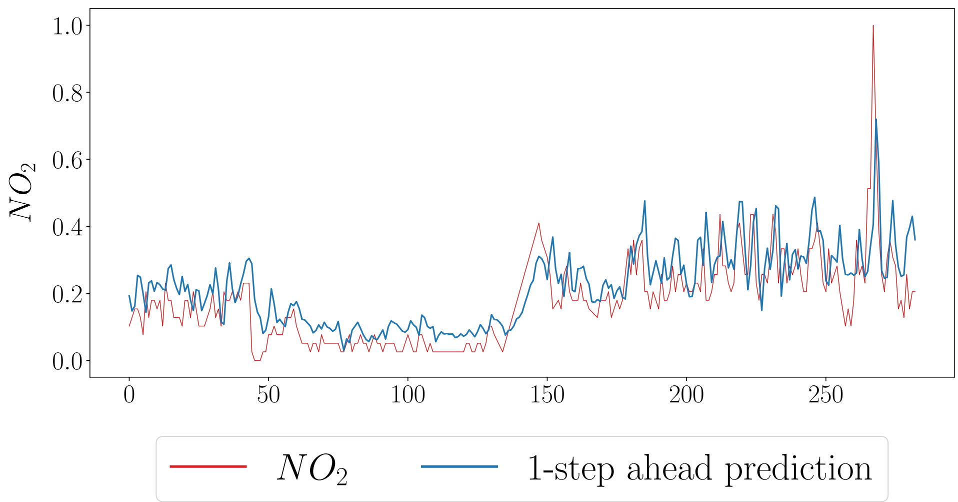


Figure 3.16: Times series of 1-step ahead predictions for NO_2 evaluated on test with the best model (RF).

Method	Number of selected attributes	$\mathcal{H}_R$	$\mathcal{H}_T$	Run time (minutes)
Generation-based fixed evolution control	14	0.0909	0.1421	25.17
O1O2 in [40]	2	0.1234	0.1442	9.80
All attributes	84	0.0560	0.2763	0.01

Table 3.9: Comparison of the best generation-based fixed evolution control model with other surrogate-assisted approach and with the forecast model with all attributes for the air quality forecast problem, sorted from best to worse evaluation of $\mathcal{H}_T$.

maintaining good model interpretability.

On the other hand, a study of the behavior of the method has been carried out depending on the update frequency of the surrogate model. Logically, the lower the frequency, the longer the method takes since the update period is shorter. Though in the indoor temperature problem there are no significant differences between the frequencies, in the air quality problem it has been established that a frequency of 50 achieves better solutions on average. Despite the increase in computational time with respect to other methods presented, it is one of the best forecasting methods of all those analyzed.

3.5 Time series classification and clustering

During the PhD stay at Siemens AG Österreich in Vienna (Austria), a study of different techniques for time series classification and clustering has been conducted. This study was carried out on a dataset containing hourly energy, heating and water measurements from several sensors inside smart buildings. These experiments have helped me to have a better understanding of time series applied in a classification context. The acquired experience was useful for the development of a method applicable to regression and classification problems [48]. The following results are being considered for release as part of a broader study within the field of smart buildings.

3.5.1 Time series classification

To apply classification methods the time series have been studied and aggregated for three granularities: day (566028 time series with a length of 24), week (97524 time series with a length of 168) and month (16175 time series with a length of 672). The series were normalized and divided into 80% for training and 20% for testing. Several time series classification techniques such as kNN, *learning shapelets* [51], *symbolic aggregate approximation - vector space model* (SAX-VSM) [52], *bag-of-SFA symbols in vector space* (BOSSVS) [53] and *TimeSeriesForest* [54] have been considered. After a preliminary study, it was decided to use SAX-VSM and TimeSeriesForest, as they were the best-performing models. The results are shown in Tables 3.10, 3.11 and 3.12.

	precision	**recall**	**f1-score**	**support**
Energy	0.69	0.73	0.71	22767
Heating	0.51	0.34	0.41	15952
Cold water	0.32	0.53	0.40	16009
Warm water	0.75	0.60	0.67	30632
accuracy			0.57	85360
macro avg	0.57	0.55	0.55	85360
weighted avg	0.61	0.57	0.58	85360

	precision	**recall**	**f1-score**	**support**
Energy	0.98	0.97	0.98	24271
Heating	0.72	0.88	0.79	8675
Cold water	0.73	0.72	0.72	26667
Warm water	0.75	0.71	0.73	25747
accuracy			0.80	85360
macro avg	0.79	0.82	0.80	85360
weighted avg	0.80	0.80	0.80	85360

Table 3.10: Classification report for SAX-VSM and TimeSeriesForest (left to right) with day granularity.

	precision	**recall**	**f1-score**	**support**
Energy	0.84	0.88	0.86	2681
Heating	0.60	0.60	0.60	1522
Cold water	0.56	0.68	0.61	2607
Warm water	0.84	0.68	0.76	3852
accuracy			0.72	10662
macro avg	0.71	0.71	0.71	10662
weighted avg	0.74	0.72	0.72	10662

	precision	**recall**	**f1-score**	**support**
Energy	0.99	0.99	0.99	2797
Heating	0.78	0.87	0.82	1359
Cold water	0.79	0.78	0.79	3242
Warm water	0.83	0.79	0.81	3264
accuracy			0.85	10662
macro avg	0.85	0.86	0.85	10662
weighted avg	0.85	0.85	0.85	10662

Table 3.11: Classification report for SAX-VSM and TimeSeriesForest (left to right) with week granularity.

	precision	**recall**	**f1-score**	**support**
Energy	0.89	0.92	0.91	794
Heating	0.65	0.76	0.70	467
Cold water	0.63	0.71	0.67	816
Warm water	0.88	0.72	0.79	1158
accuracy			0.77	3235
macro avg	0.76	0.78	0.77	3235
weighted avg	0.79	0.77	0.78	3235

	precision	**recall**	**f1-score**	**support**
Energy	0.99	0.98	0.98	825
Heating	0.82	0.87	0.84	521
Cold water	0.78	0.76	0.77	944
Warm water	0.82	0.81	0.81	945
accuracy			0.85	3235
macro avg	0.85	0.86	0.85	3235
weighted avg	0.85	0.85	0.85	3235

Table 3.12: Classification report for SAX-VSM and TimeSeriesForest (left to right) with month granularity.

For both algorithms, the classification between cold and warm water is the worst, as they are time series that follow a similar pattern. Energy consumption shows the best classification, since it follows a fairly characteristic pattern compared to the rest of the time series. The case with granularity of one day is the one with the worst accuracy, this is due to the fact that the length of the time series is short and it is difficult to find distinguishable characteristics between the different series. The performance of the models with one week and one month granularity are very similar.

3.5.2 Time series clustering

In order to check if there is any pattern within the energy consumption time series, these series have been grouped by ID. Therefore, there are 133 time series in total. For each of the 7 days of the week, each of its 24 hours has been averaged by hour. That is, the average of the values of all Mondays at 12 pm, at 1 am, and so on. Those new time series have a length of 168 (7 days $\times$ 24 hours). To establish the optimal number of clusters and their consistency, the *silhouette index* [55] of the time series set is calculated. Figure 3.17 shows that 3 clusters are a good size since there are hardly any misclassified samples.

The *k-means* [56] clustering algorithm has been applied to group time series with similar patterns. Figure 3.18 shows the result, the red line represents the mean value of all the time series in the cluster in question. Three different patterns of behavior have been established. Cluster 0 would belong to houses with workers, and clusters 1 and 3 would belong to houses where some inhabitant does not work.

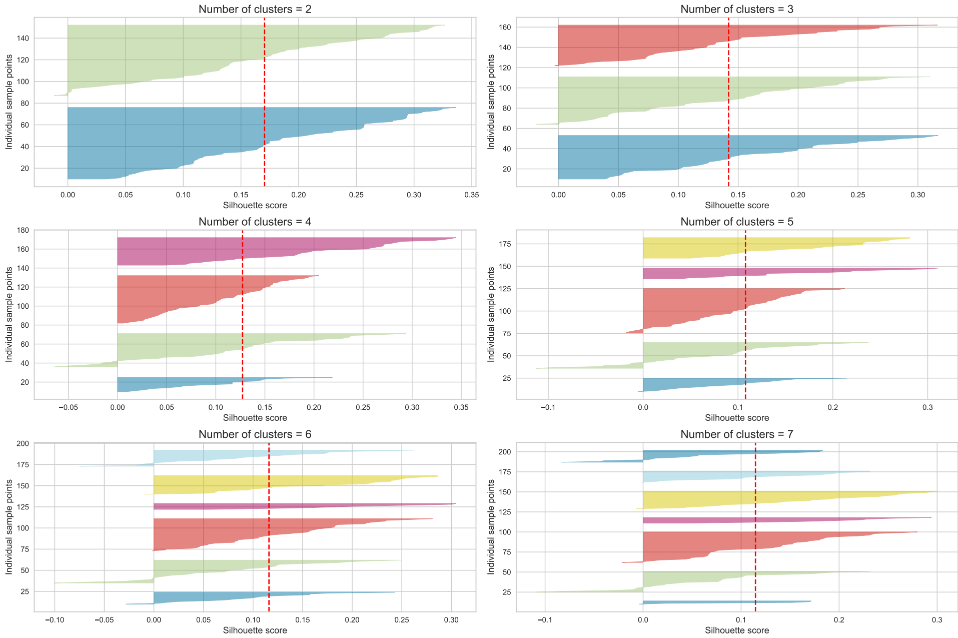


Figure 3.17: Silhouette score for 2 to 7 clusters for averaged hour of days of the week.

Generally, the highest consumption peaks occur in the evening hours around 6 pm, this fits the time when workers go home. In contrast, the hours with the lowest energy consumption are between 12 am and 6 am, an interval that coincides with sleeping hours. On weekends, a general increase in consumption is observed, since it is likely that all the inhabitants of the house are at home.

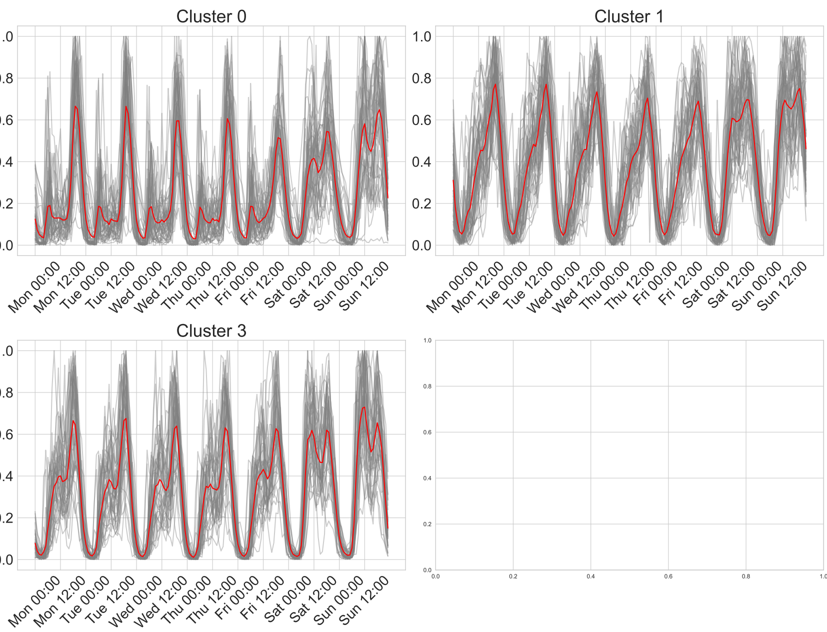


Figure 3.18: 3 clusters k-means for energy IDs with mean per hour of days of the week.

Chapter 4

Conclusions and future works

In this last chapter, the conclusions of the work carried out and possible future works are presented.

4.1 Conclusions

In recent years, the amount of information available has increased dramatically, partly due to the proliferation of new technologies and the way in which data is collected. Working with a large amount of data is in many cases unfeasible and makes it difficult to build comprehensible learning models. For this reason, FS techniques are very useful as they reduce the complexity of the available information, especially in the context of time series, where redundant information may be present.

This thesis develops several FS techniques for deep learning based on surrogate-assisted MOEAs applied to time series forecasting. First, a previous study of the behavior of different machine learning and deep learning techniques in a time series problem has been carried out. In addition, a multi-objective evolutionary algorithm in conjunction with ensemble learning has been proposed in order to find prediction models with spatio-temporal characteristics of a prediction problem where no data are available. Besides, their contribution to the forecasts has been analyzed. After this, the FS problem has been formalized as different multi-objective optimization problems, in which the behavior of the objective functions and their contribution to the selection of optimal features has been studied. The performance of various MOEAs such as NSGA-II, NSGA-III, SPEA2, MOEA/D, IBEA, ϵ-MOEA or ϵ-NSGA-II as well as their best parameters have also been analyzed. To reduce the computational cost of a wrapper-type FS method with deep learning, three different approaches for FS with deep learning based on surrogate-assisted multi-objective evolutionary algorithms have been proposed and validated. The developed methods have been applied mainly in air quality time series. The techniques proposed in this thesis have been compared with other existing FS methods such as CancelOut, real wrappers, hybrid filter-wrapper or RF.

The main conclusions drawn from this thesis after the execution of all the experiments are as follows:

- The adoption of a complete methodology for the evaluation and comparison of learning algorithms has allowed to obtain unified and adapted results in order to solve any prediction problem with time series.
- Recurrent neural networks, such as LSTM and GRU, have been able to capture the complexity of time series and build accurate and reliable predictive models. Among the analyzed machine learning techniques, RF has presented a satisfactory performance when applied to time series forecasting.
- A multi-criteria decision-making process has allowed to pool several performance metrics and to establish a more appropriate comparison between different learning algorithms in the context of time series forecasting problems.
- For air quality forecasting with time series in an area for which no information is available, the prediction has been approximated with multi-objective evolutionary algorithms using forecasts from other geographically nearby areas.
- Surrogate-assisted multi-objective evolutionary algorithms has allowed feature selection in expensive problems such as time series forecasting based on deep learning. Additionally, dimensionality reduction has simplified the predictive models built thus increasing their interpretability and, helping to stop perceiving the models as a "black box".
- The use of a surrogate-assisted MOEAs with a deep learning algorithm for feature selection has managed to find a satisfactory subset of features in a shorter computational time compared to a conventional wrapper-type feature selection method.
- Among all the MOEAs studied, NSGA-II is the one that has obtained the best results in terms of hypervolume, compared to other MOEAs of the state of the art.
- Generation-based fixed evolution control approach allows information to be efficiently added to surrogate models within the feature selection process. While this succeeds in finding better subsets of attributes and improving prediction results, it does so at the cost of increasing the computational time of the process, since after a fixed number of evaluations of the evolutionary algorithm the surrogate model has to be retrained and/or incremented. Although, it is more efficient than conventional wrapper FS methods.
- Prediction models have been identified in various real contexts (Poland, Murcia, Valencia) that potentially allow forecasting in the near future and that can help institutions to make decisions on environmental issues. Another significant factor of the proposed methods in the environmental field is that by having a shorter computational time, the carbon footprint can be reduced, contributing to the *European Green Deal*[1]. In the social field, the proposed methods contribute to *Explainable Artificial Intelligence* (XIA) [57], and are aligned with the objectives of initiatives

[1] `https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/european-green-deal_en`

such as the *White Paper on Artificial Intelligence*[2] of the European Commission and Spain's *National Artificial Intelligence Strategy*[3].

4.2 Future works

After the study described in this thesis, the following lines of research to be developed in the future are considered:

- Include the feature selection process within the spatio-temporal approach with LR, adding a new objective within the optimization problem that tries to minimize the number of attributes selected in LR. This application in other more complex learning algorithms, such as neural networks, will also be studied.
- Apply the surrogate-assisted and multi-surrogate assisted multi-objective evolutionary algorithm feature selection methods for forecasting other air quality related time series such as CO_2, $PM_{2.5}$ or PM_{10} and compare its performance with the results from NO_2.
- Use other deep learning algorithms such as GRU or QRNN as a surrogate model within the evolutionary multi-objective feature selection method and compare their performance with the current method with LSTM. Additionally, the use of other multi-step ahead forecasting strategies, such as direct multi-step forecast or direct-recursive hybrid, will be analyzed.
- MOEAs have been successfully used in the search for the optimal architecture in predictive models, especially in deep learning. In the development of this thesis, it has been shown that the use of MOEAs for feature selection with a LSTM-based model obtains good results. However, there are no works at present that combine architecture search with attribute selection, so it is a very interesting field of study and will be addressed in future research, both for regression and classification problems.
- Apply the proposed feature selection methods in other fields such as image recognition and natural language processing.

[2] `https://commission.europa.eu/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en`

[3] `https://portal.mineco.gob.es/es-es/ministerio/areas-prioritarias/Paginas/inteligencia-artificial.aspx`

Chapter 5

Publications composing the doctoral thesis

5.1 A time series forecasting based multi-criteria methodology for air quality prediction

Abstract There is a very extensive literature on the design and test of models of environmental pollution, especially in the atmosphere. Current and recent models, however, are focused on explaining the causes and their temporal relationships, but do not explore, in full detail, the performances of pure forecasting models. We consider here three years of data that contain hourly nitrogen oxides concentrations in the air; exposure to high concentrations of these pollutants has been indicated as potential cause of numerous respiratory, circulatory, and even nervous diseases. Nitrogen oxides concentrations are paired with meteorological and vehicle traffic data for each measure. We propose a methodology based on exactness and robustness criteria to compare different pollutant forecasting models and their characteristics. 1DCNN, GRU and LSTM deep learning models, along with Random Forest, Lasso Regression and Support Vector Machines regression models, are analyzed with different window sizes. As a result, our best models offer a 24-hours ahead, very reliable prediction of the concentration of pollutants in the air in the considered area, which can be used to plan, and implement, different kinds of interventions and measures to mitigate the effects on the population.

Title	A time series forecasting based multi-criteria methodology for air quality prediction
Authors	Raquel Espinosa, José Palma, Fernando Jiménez, Joanna Kamińska, Guido Sciavicco, Estrella Lucena-Sánchez
Journal	Applied Soft Computing
Impact factor (2021)	8.263
JCR Rank (2021)	Computer science, interdisciplinary applications: 11/112 (D1) Computer science, artificial intelligence: 23/145 (Q1)
Cited by	21 (Google Scholar)
Publisher	Elsevier
Date	7 September 2021
ISSN	1568-4946
DOI	`https://doi.org/10.1016/j.asoc.2021.107850`
State	Published
Contribution	Term, conceptualization, methodology, software, validation, investigation, resources, data curation, writing – original draft, writing – review and editing, visualization, project administration

5.2 Multi-objective evolutionary spatio-temporal forecasting of air pollution

Abstract Nowadays, air pollution forecasting modeling is vital to achieve an increase in air quality, allowing an improvement of ecosystems and human health. It is important to consider the spatial characteristics of the data, as they allow us to infer predictions in those areas for which no information is available. In the current literature, there are a large number of proposals for spatio-temporal air pollution forecasting. In this paper we propose a novel spatio-temporal approach based on multi-objective evolutionary algorithms for the identification of multiple non-dominated linear regression models and their combination in an ensemble learning model for air pollution forecasting. The ability of multi-objective evolutionary algorithms to find a Pareto front of solutions is used to build multiple forecast models geographically distributed in the area of interest. The proposed method has been applied for one-week NO_2 prediction in southeastern Spain and has obtained promising results in statistical comparison with other approaches such as the union of datasets or the interpolation of the predictions for each monitoring station. The validity of the proposed spatio-temporal approach is thus demonstrated, opening up a new field in air pollution engineering.

Title	Multi-objective evolutionary spatio-temporal forecasting of air pollution
Authors	Raquel Espinosa, Fernando Jiménez, José Palma
Journal	Future Generation Computer Systems
Impact factor (2021)[4]	7.307
JCR Rank (2021)[4]	Computer science, theory & methods: 10/110 (D1)
Cited by	4 (Google Scholar)
Publisher	Elsevier
Date	31 May 2022
ISSN	0167-739X
DOI	`https://doi.org/10.1016/j.future.2022.05.020`
State	Published
Contribution	Conception and design of the study, acquisition of data, analysis and interpretation of data, point by point revision of reviewer's comments

[4]At the time of publication of this thesis, the impact factor and JCR rank data for the year 2022 were not available.

5.3 Multi-surrogate assisted multi-objective evolutionary algorithms for feature selection in regression and classification problems with time series data

Abstract Feature selection wrapper methods are powerful mechanisms for reducing the complexity of prediction models while preserving and even improving their precision. Meta-heuristic methods, such as multi-objective evolutionary algorithms, are commonly used as search strategies in feature selection wrapper methods since they allow minimizing the cardinality of the attribute subset and simultaneously maximizing the predictive capacity of the model. However, in high-dimensional problems, multi-objective evolutionary algorithms for wrapper-type feature selection may require excessive computational time, sometimes impractical, especially when the learning algorithm has a high computational cost, such as deep learning. To address this drawback, in this paper we propose a multi-surrogate assisted multi-objective evolutionary algorithm for feature selection, specially designed to improve generalization error. The proposed method has been compared with conventional feature selection wrapper methods that use random forest, support vector machine and long short-term memory learning algorithms to evaluate subsets of attributes. The experiments have been carried out with regression and classification problems with time series data for air quality forecasting in the south-east of Spain and for indoor temperature forecasting in a domotic house. The results demonstrate the superiority of the proposed multi-surrogate assisted method over conventional wrapper methods using the same run times.

Title	Multi-surrogate assisted multi-objective evolutionary algorithms for feature selection in regression and classification problems with time series data
Authors	Raquel Espinosa, Fernando Jiménez, José Palma
Journal	Information Sciences
Impact factor (2021)[5]	8.233
JCR Rank (2021)[5]	Computer science, information systems: 16/164 (D1)
Cited by	4 (Google Scholar)
Publisher	Elsevier
Date	10 December 2022
ISSN	0020-0255
DOI	`https://doi.org/10.1016/j.ins.2022.12.004`
State	Published
Contribution	Project administration, conceptualization, methodology, data curation, visualization, investigation, software, writing

[5] At the time of publication of this thesis, the impact factor and JCR rank data for the year 2022 were not available.

5.4 Surrogate-assisted and filter-based multiobjective evolutionary feature selection for deep learning

Abstract Feature selection 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 paper, we propose new attribute subset evaluation feature selection methods for deep learning of the wrapper, filter and wrapper-filter hybrid types, where multi-objective 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 feature selection techniques used in the literature.

Title	Surrogate-assisted and filter-based multi-objective evolutionary feature selection for deep learning
Authors	Raquel Espinosa, Fernando Jiménez, José Palma
Journal	IEEE Transactions on Neural Networks and Learning Systems
Impact factor (2021)[6]	14.255
JCR Rank (2021)[6]	Computer science, artificial intelligence: 6/145 (D1) Computer science, theory & methods: 4/110 (D1) Computer science, hardware & architecture: 1/54 (D1)
Publisher	IEEE
Date	12 January 2023
ISSN	2162-237X
DOI	`https://doi.org/10.1109/TNNLS.2023.3234629`
State	Published
Contribution	Conceptualization, methodology, visualization, investigation, software, writing, point by point revision of reviewer's comments

[6]At the time of publication of this thesis, the impact factor and JCR rank data for the year 2023 were not available.

Bibliography

[1] B. M. Henrique, V. A. Sobreiro, H. Kimura, Literature review: Machine learning techniques applied to financial market prediction, Expert Systems with Applications 124 (2019) 226–251.

[2] A. A. Cook, G. Mısırlı, Z. Fan, Anomaly detection for IoT time-series data: A survey, IEEE Internet of Things Journal 7 (7) (2019) 6481–6494.

[3] S. Ballı, Data analysis of Covid-19 pandemic and short-term cumulative case forecasting using machine learning time series methods, Chaos, Solitons & Fractals 142 (2021) 110512.

[4] P. R. Srivastava, Z. J. Zhang, P. Eachempati, Deep neural network and time series approach for finance systems: predicting the movement of the Indian stock market, Journal of Organizational and End User Computing (JOEUC) 33 (5) (2021) 204–226.

[5] C.-Y. Lin, Y.-S. Chang, S. Abimannan, Ensemble multifeatured deep learning models for air quality forecasting, Atmospheric Pollution Research 12 (5) (2021) 101045.

[6] M. Khodayar, M. E. Khodayar, S. M. J. Jalali, Deep learning for pattern recognition of photovoltaic energy generation, The Electricity Journal 34 (1) (2021) 106882.

[7] S. Sharma, K. K. Bhatt, R. Chabra, N. Aneja, A Comparative Performance Model of Machine Learning Classifiers on Time Series Prediction for Weather Forecasting, in: Advances in Information Communication Technology and Computing, Springer, 2022, pp. 577–587.

[8] R. Bellman, Dynamic programming, Science 153 (3731) (1966) 34–37.

[9] M. Verleysen, D. François, The curse of dimensionality in data mining and time series prediction, in: J. Cabestany, A. Prieto, F. Sandoval (Eds.), Computational Intelligence and Bioinspired Systems, Springer Berlin Heidelberg, Berlin, Heidelberg, 2005, pp. 758–770.

[10] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, Journal of machine learning research 3 (Mar) (2003) 1157–1182.

[11] K. Deb, Multi-objective optimisation using evolutionary algorithms: an introduction, in: Multi-objective evolutionary optimisation for product design and manufacturing, Springer, 2011, pp. 3–34.

[12] Y. Xue, Y. Tang, X. Xu, J. Liang, F. Neri, Multi-objective feature selection with missing data in classification, IEEE Transactions on Emerging Topics in Computational Intelligence 6 (2) (2021) 355–364.

[13] Y. Xue, X. Cai, F. Neri, A multi-objective evolutionary algorithm with interval based initialization and self-adaptive crossover operator for large-scale feature selection in classification, Applied Soft Computing 127 (2022) 109420.

[14] Y. Jin, Surrogate-assisted evolutionary computation: Recent advances and future challenges, Swarm and Evolutionary Computation 1 (2) (2011) 61–70.

[15] J. Li, P. Wang, H. Dong, J. Shen, C. Chen, A classification surrogate-assisted multi-objective evolutionary algorithm for expensive optimization, Knowledge-Based Systems 242 (2022) 108416.

[16] F. Li, L. Gao, W. Shen, Surrogate-Assisted Multi-Objective Evolutionary Optimization With Pareto Front Model-Based Local Search Method, IEEE Transactions on Cybernetics (2022).

[17] Air Pollution Data Portal (2022).
URL `https://www.who.int/news/item/04-04-2022-billions-of-people-still-breathe-unhealthy-air-new-who-data`

[18] G.-P. Bălă, R.-M. Râjnoveanu, E. Tudorache, R. Motișan, C. Oancea, Air pollution exposure—the (in) visible risk factor for respiratory diseases, Environmental Science and Pollution Research 28 (16) (2021) 19615–19628.

[19] J. L. Domingo, J. Rovira, Effects of air pollutants on the transmission and severity of respiratory viral infections, Environmental research 187 (2020) 109650.

[20] D. A. Glencross, T.-R. Ho, N. Camina, C. M. Hawrylowicz, P. E. Pfeffer, Air pollution and its effects on the immune system, Free Radical Biology and Medicine 151 (2020) 56–68.

[21] Health impacts of air pollution in Europe (2022).
URL `https://www.eea.europa.eu/publications/air-quality-in-europe-2022/health-impacts-of-air-pollution`

[22] R. M. De Carvalho, C. F. Szlafsztein, Urban vegetation loss and ecosystem services: The influence on climate regulation and noise and air pollution, Environmental Pollution 245 (2019) 844–852.

[23] D. Dua, C. Graff, UCI machine learning repository (2017).
URL `http://archive.ics.uci.edu/ml`

[24] F. Zamora-Martinez, P. Romeu, P. Botella-Rocamora, J. Pardo, On-line learning of indoor temperature forecasting models towards energy efficiency, Energy and Buildings 83 (2014) 162–172.

[25] R. Espinosa, J. Palma, F. Jiménez, J. Kamińska, G. Sciavicco, E. Lucena-Sánchez, A time series forecasting based multi-criteria methodology for air quality prediction, Applied Soft Computing 113 (2021) 107850.

[26] S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, D. J. Inman, 1D convolutional neural networks and applications: A survey, Mechanical systems and signal processing 151 (2021) 107398.

[27] K. Cho, B. Van Merriënboer, D. Bahdanau, Y. Bengio, On the properties of neural machine translation: Encoder-decoder approaches, arXiv preprint arXiv:1409.1259 (2014).

[28] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (8) (1997) 1735–1780.

[29] L. Breiman, Random forests, Machine learning 45 (1) (2001) 5–32.

[30] R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society: Series B (Methodological) 58 (1) (1996) 267–288.

[31] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, B. Scholkopf, Support vector machines, IEEE Intelligent Systems and their applications 13 (4) (1998) 18–28.

[32] R. Espinosa, F. Jiménez, J. Palma, Multi-objective evolutionary spatio-temporal forecasting of air pollution, Future Generation Computer Systems 136 (2022) 15–33.

[33] K. Deb, A. Pratap, S. Agarwal, T. Meyarivan, A fast and elitist multiobjective genetic algorithm: NSGA-II, IEEE transactions on evolutionary computation 6 (2) (2002) 182–197.

[34] Q. Zhang, H. Li, MOEA/D: A multiobjective evolutionary algorithm based on decomposition, IEEE Transactions on evolutionary computation 11 (6) (2007) 712–731.

[35] E. Zitzler, M. Laumanns, L. Thiele, SPEA2: Improving the strength Pareto evolutionary algorithm, TIK-report 103 (2001).

[36] J. Bradbury, S. Merity, C. Xiong, R. Socher, Quasi-recurrent neural networks, arXiv preprint arXiv:1611.01576 (2016).

[37] R. Vang-Mata, Multilayer Perceptrons: Theory and Applications, Nova Science Publishers, 2020.

[38] D. W. Aha, D. Kibler, M. K. Albert, Instance-based learning algorithms, Machine learning 6 (1) (1991) 37–66.

[39] J. Li, A Review of Spatial Interpolation Methods for Environmental Scientists, Geoscience Australia, Record 2008/23, 2008.

[40] R. Espinosa, F. Jiménez, J. Palma, Surrogate-assisted and filter-based multi-objective evolutionary feature selection for deep learning, IEEE Transactions on Neural Networks and Learning Systems (2023) 1–15.

[41] M. A. Hall, Correlation-based feature selection for machine learning, Ph.D. thesis, The University of Waikato (1999).

[42] I. Kononenko, E. Šimec, M. Robnik-Šikonja, Overcoming the myopia of inductive learning algorithms with RELIEFF, Applied Intelligence 7 (1) (1997) 39–55.

[43] K. Deb, H. Jain, An evolutionary many-objective optimization algorithm using reference-point-based nondominated sorting approach, part I: solving problems with box constraints, IEEE transactions on evolutionary computation 18 (4) (2013) 577–601.

[44] E. Zitzler, S. Künzli, Indicator-based selection in multiobjective search, in: International conference on parallel problem solving from nature, Springer, 2004, pp. 832–842.

[45] Z. Fan, W. Li, X. Cai, H. Huang, Y. Fang, Y. You, J. Mo, C. Wei, E. Goodman, An improved epsilon constraint-handling method in MOEA/D for CMOPs with large infeasible regions, Soft Computing 23 (23) (2019) 12491–12510.

[46] J. B. Kollat, P. M. Reed, The value of online adaptive search: A performance comparison of NSGAII, ε-NSGAII and εMOEA, in: International Conference on Evolutionary Multi-Criterion Optimization, Springer, 2005, pp. 386–398.

[47] V. Borisov, J. Haug, G. Kasneci, Cancelout: A layer for feature selection in deep neural networks, in: International conference on artificial neural networks, Springer, 2019, pp. 72–83.

[48] R. Espinosa, F. Jiménez, J. Palma, Multi-surrogate assisted multi-objective evolutionary algorithms for feature selection in regression and classification problems with time series data, Information Sciences 622 (2023) 1064–1091.

[49] A. Gepperth, B. Hammer, Incremental learning algorithms and applications, in: European symposium on artificial neural networks (ESANN), 2016.

[50] F. X. Diebold, R. S. Mariano, Comparing predictive accuracy, Journal of Business & economic statistics 20 (1) (2002) 134–144.

[51] J. Grabocka, N. Schilling, M. Wistuba, L. Schmidt-Thieme, Learning time-series shapelets, in: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 2014, pp. 392–401.

[52] P. Senin, S. Malinchik, SAX-VSM: Interpretable time series classification using sax and vector space model, in: 2013 IEEE 13th international conference on data mining, IEEE, 2013, pp. 1175–1180.

[53] P. Schäfer, Scalable time series classification, Data Mining and Knowledge Discovery 30 (5) (2016) 1273–1298.

[54] H. Deng, G. Runger, E. Tuv, M. Vladimir, A time series forest for classification and feature extraction, Information Sciences 239 (2013) 142–153.

[55] A. Starczewski, A. Krzyżak, Performance evaluation of the silhouette index, in: International conference on artificial intelligence and soft computing, Springer, 2015, pp. 49–58.

[56] H.-H. Bock, Clustering methods: a history of k-means algorithms, Selected contributions in data analysis and classification (2007) 161–172.

[57] U. Kamath, J. Liu, Explainable artificial intelligence: An introduction to interpretable machine learning, Springer, 2021.

Appendix A

Abbreviations

Abbreviation	Meaning
1D-CNN	1 Dimension Convolutional Neural Network
BOSSVS	Bag-Of-SFA Symbols in Vector Space
CC	Correlation Coefficient
CO_2	Carbon Dioxide
FS	Feature Selection
GRU	Gated Recurrent Unit
IBEA	Indicator-Based Evolutionary Algorithm
IDW	Inverse Distance Weighting
KNN	K Nearest Neighbors
LR	Linear Regression
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MLP	Multi-layer Perceptron
MOEA	Multi-Objective Evolutionary Algorithm
MOEA/D	Multi-Objective Evolutionary Algorithm based on Decomposition
NO	Nitrogen Monoxide
NO_2	Nitrogen Dioxide
NO_X	Nitrogen Oxides
NSGA-II	Non-dominated Sorting Genetic Algorithm II
NSGA-III	Non-dominated Sorting Genetic Algorithm III
O_3	Ozone
PM	Particulate Matter
QRNN	Quasi-Recurrent Neural Network
RF	Random Forest
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Network
SAX-VSM	Symbolic Aggregate Approximatioon - Vector Space Model
SPEA2	Strength Pareto Evolutionary Algorithm 2
SVM	Support Vector Machine
TMP	Temperature
UCI	University of California Irvine

Table A.1: Abbreviations.