Information Fusion 74 (2021) 1–16

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Artificial intelligence and healthcare: Forecasting of medical bookings through multi-source time-series fusion

Francesco Piccialli [a], Fabio Giampaolo [a,*], Edoardo Prezioso [b], David Camacho [c], Giovanni Acampora [d]

[a] Department of Mathematics and Applications "Renato Caccioppoli", University of Naples Federico II, Italy
[b] CINI - Consorzio Nazionale Interuniversitario per l'Informatica, ITEM-SAVY Research Laboratory, Italy
[c] Universidad Politecnica de Madrid, AIDA - Applied Intelligence and Data Analysis Research Group, Spain
[d] Department of Physics "Ettore Pancini", University of Naples Federico II, Italy

ARTICLE INFO

Keywords:
Artificial intelligence
Deep Learning
Multi-source time-series
Healthcare

ABSTRACT

Nowadays, Artificial intelligence (AI), combined with the digitalization of healthcare, can lead to substantial improvements in Patient Care, Disease Management, Hospital Administration, and supply chain effectiveness. Among predictive analytics tools, time series forecasting represents a central task to support healthcare management in terms of bookings and medical services predictions. In this context, the development of flexible frameworks to provide robust and reliable predictions became a central point in this healthcare innovation process. This paper presents and discusses a multi-source time series fusion and forecasting framework relying on Deep Learning. By combining weather, air-quality and medical bookings time series through a feature compression stage which preserves temporal patterns, the prediction is provided through a flexible ensemble technique based on machine learning models and a hybrid neural network. The proposed system is able to predict the number of bookings related to a specific medical examination for a 7-days horizon period. To assess the proposed approach's effectiveness, we rely on time series extracted from a real dataset of administrative e-health records provided by the Campania Region health department, in Italy.

1. Introduction

Nowadays, the predictive analytics application field has a global value of more than 10 billion dollars. In a hyperconnected world where tons of data are generated continuously, it probably represents one of the most significant components for business and research activities, since it allows to describe real dynamics and provide future predictions.

Healthcare is one of the contexts that can benefit most from predictive analytics. For example (i) by providing tools to improve care by foreseeing the possibility of an event or (ii) by considering a patient to be hospitalized, thereby allowing the healthcare system to offer tailored treatments. Moreover, these tools are proven to be useful in the continuation of care, in the monitoring of pathologies, in preventing chronic diseases or also in better understanding biological processes [1].

The extraction, the real understanding, and the accurate analysis and definition of the value of patient-related data pools will create a whole "new value" in hospitals, guaranteeing more accurate services to citizens/patients and distributing resources in a more balanced and virtuous way. Especially in the last few years, Artificial Intelligence (AI) and Machine Learning (ML), thanks to the increasing of computational power and the design of efficient algorithms [2–4], have become increasingly popular [5] in this renewal process of healthcare since they can improve the ability to analyse complex data patterns related to this context [6]: compared to business intelligence carried out with traditional technologies, AI allows to better estimate and understands current trends, perform predictive analysis, and find correlations between different phenomena.

This increased analytical capacity can be very useful in resource planning, for example, in the management of contacts (i.e. contact centres), visits reservations, beds and operating rooms allocation. Additional information and awareness on resource requests' dynamics allow us to articulate and segment the services' offer to maximize the results obtained.

With the use of AI technologies, it is possible to create optimizing tools to plan and schedule resources that, for example, suggest to stakeholders the settings of a reservation system agendas or how to manage queues in a contact centre [7]. In this context, predictive

* Corresponding author.
E-mail addresses: francesco.piccialli@unina.it (F. Piccialli), fabio.giampaolo@consorzio-cini.it (F. Giampaolo), edoardo.prezioso@consorzio-cini.it (E. Prezioso), david.camacho@upm.es (D. Camacho), giovanni.acampora@unina.it (G. Acampora).

https://doi.org/10.1016/j.inffus.2021.03.004
Received 20 October 2020; Received in revised form 2 February 2021; Accepted 21 March 2021
Available online 27 March 2021

analytics have also been exploited to predict the patients who fail to attend their appointments [8], since this phenomenology usually disrupts medical management and it also resources consuming. Moreover, it has been applied to predict the scheduled hospital attendance [9] or to predict chronic disease's hospitalizations since it also can optimize the scheduling process of a medical structure [10] reducing hospital care costs [11].

With these premises, the need for methodologies more and more capable of reliable predictions emerges as an essential requirement to develop new optimizing and cost-effective strategies, especially in the healthcare domain. In particular, time series forecasting results of primary importance to improve scheduling procedures and staff management. In this scenario, this paper proposes a framework for reliable multi-source medical structures workload and visit reservations time series prediction using a hybrid neural network forecast combiner and data fusion. The idea behind the work is to provide a workflow capable of exploiting short-to-mid relations between the time series under consideration, in the case of *respiratory diseases bookings* time series provided by the healthcare authorities of Campania Region (Italy) and other related time series, i.e. *air-quality* and *weather* sequences in this work, also themselves connected.

The main contribution of this paper can be summarized as follows:

- A prediction workflow for reliable short-term forecasting of workload/reservations in the healthcare context exploiting multi-source time sequences.
- A data fusion strategy to combine multi-source temporal data, which compresses the feature set in a lower-dimensional space preserving the features themselves' temporal relations.
- An ensemble methodology of different Machine Learning models through a hybrid neural network which produces a robust 7-days forecast of daily time series, in this specific context the sequence of *respiratory diseases bookings*.

An extended comparison with both other forecasting models and ensemble strategies shows that the proposed workflow enhances the prediction's robustness, often resulting in improved accuracy, as evidenced by the promising results with respect to the MAE and $RMSE$ error metrics, and to the R^2 coefficient.

The rest of the paper is organized as follows: Section 2 discusses the background of this work and motivate architectural choices of the framework; Section 3 presents a correlation and causality analysis among the considered time series, whose aim is to validate the assumption on mutual relations between them; Section 4 describes the structure of the framework; Section 5 presents and discusses the experimental results. Finally, Section 6 concludes the paper and offers an overview of further studies about the topic.

2. Background

The exponential growth of information availability has made the necessity of extracting useful knowledge from large batches of data more and more compelling. Moreover, aggregating information coming from different sources has opened new scenarios in complex modelling phenomena. In this context, Machine Learning and Deep Learning methodologies have been successfully applied in describing such relations, and today they are adopted as the state-of-the-art approach in understanding underlying data patterns and in providing predictive analytics. This second task, in particular, has become the context in which Supervised Learning techniques are mostly applied since they have been proven to afford reliable and flexible results. Among predictive tasks, the prediction of temporal sequences is one of the most challenging problems and, at the same time, one of the most faced since its importance in decision-making for organizations and management.

2.1. Time series forecasting

Over the years, many reliable strategies for Time Series prediction have been explored: during the 50 s and 60 s, starting from the simplistic idea that a future observation linearly depends on past values of the sequence, some extensions to this concept have been proposed, as the Exponential Smoothing suggested by Holt [12]. In the 70 s, thanks to Box and Jenkins [13], the first attempt in modelling autocorrelation among data themselves paved the way for the class of autoregressive integrated moving average methods. These models are still widely used today with some modifications to consider the more complex behaviours of a time sequence, i.e. seasonalities and the dependence from exogenous variables. The advent of algorithms capable of modelling relations among available information without imposing any constraint on the relationship itself has made available new ways in facing the forecasting problem. In particular, thanks to the inherent ability to extract complex patterns directly from data, artificial neural networks have successfully proven their capability in predictive analysis of temporal sequences during the last twenty years. Based on different approaches [14–16], many different neural network strategies have been used in time series prediction problems [17–21] thanks to the flexibility given by the possibility of combining different neural structures to exploit different aspects of data.

2.2. Ensemble techniques in time series prediction

Thanks to the increasing computational power, The research on methodologies that reduce the uncertainty connected to the usage of a single model, although highly optimized, has become more and more intense in the last years. Ensemble techniques have become a central topic in the forecasting field: finding a strategy to combine different predictions on the same temporal sequence, each of them able to exploit different aspects of the data, allows to obtain robust prediction frameworks with improved accuracy. In recent literature, ensembles for time series prediction are often based on linear combination with a weight refinement procedure [22–25], or on more complex strategies to measure the contribution of the single models as statistical methodologies [26,27], genetic algorithms [28,29] and particle swarm optimization strategies [30]. Promising results in modelling non-linear relations and complex patterns have been obtained through ensembles of neural network models, as multi-input multi-channel architectures combined utilizing common final dense layers [31–33]. However, simple algebraic combinations, as the equally weighted combination or the median, remain effective and have been proven competitive, sometimes superior, to more complex combination strategies [34,35].

2.3. Data fusion in healthcare

The idea of merging information coming from different sources to develop a comprehensive mathematical model can be found in literature since the 60 s. This concept has been applied to a wide range of research topics during the years, from pattern recognition [36] to metrology [37]. Nowadays, thanks to the huge presence of interconnected sensors and the possibility of storing a massive amount of data, data fusion is experiencing a new wave of interest also thanks to Deep Learning methodologies, whose modelling abilities of complex patterns and relations among data of different types, have made possible advanced analysis and predictive tools in many sciences fields [38–41]. In the specific context of time series forecasting, data fusion strategies often result in improved models: embedding information from different and related sources allows to exploit different aspects of the same problem by understanding their relations and producing more reliable and robust previsions [42–44] than standard univariate and multivariate models.

As regards general healthcare scenarios, strategies to combine clinical and non-clinical information to predict future health-related outcomes are widely present in literature, often in connection with DL

techniques for the prediction stage. In [45,46] two methodologies in enhancing the recognition of patients' activities based on a Kernel PCA fusion technique and RNN-based architecture for sequence modelling, and on a swarm search feature selection procedure for a machine learning framework operating on fused data, respectively, are proposed. Strategies in remote monitoring of health conditions through wearable and environmental sensors are presented in [47], where data are fused by the means of a weighting algorithm. A decision-make phase is addressed by a Fuzzy inference system in [48] where data are collected together with a multimodal algorithm, or in [49], where the feature extraction phase on the combined dataset is performed through an evolutive algorithm (SSA optimization based feature selection), and the classification outcome is obtained by a modified Deep Belief Network. In [50] structured and unstructured data in EHR are encoded through a CNN-based and an LSTM-based Neural network, and a patient vectorial representation is then used for the classification task. In [51] sensors and EHR data are combined through Information Gain and Conditional Probability approaches; a Deep learning Classifier is then trained on such data to provide heart disease prediction. In [52] biosignals and medical images are combined, and the feature selection phase is performed by a group Lasso regularization technique in order to capture temporal consequentiality of sparse data and to predict chronic disease progression. In [53–55] EHR data and medical images (MRI, images of blood cells, respectively) are fused through CNN-based models to extract the significant features and to then predict the disease progression.

Moreover, in the context we are discussing, forecasting models are often related to the prediction of the demand of a specific service, as in [56] where school zones and census tracts information are exploited to frame the forecast of vaccine delinquency and prediction of the demand of mobile clinics as non-convex optimization problems, or to predict the spreading of an illness: in [57,58] influenza or influenza-like illnesses predictions are enhanced with EHR, social media activities, internet search trends and medical surveillance data through multi-linear methods relying on different combination strategies (Autoregressive Likelihood Ratio and a Greedy optimization-based technique respectively); in [59] for the same purpose, and on a similar aggregated dataset, a strategy to combine SIR-based and statistical models by a MCMC method is presented; in [60] multiple textual data sources are exploited in an algorithmic framework which provides rare disease outbreak forecasts obtained as a weighted combination of predictions on each data source.

Until now, different strategies for both assembling data sources and obtaining robust forecasts have been mentioned. To summarize, often data fusion and feature selection are performed by weighting procedures, statistical strategies or evolutive algorithms. However, in the case complex relations exist between the features and the targets and between the features themselves, all the strategies mentioned above may not properly model patterns in a lower-dimensional representation of the original data. In the case Neural Network, or in general supervised learning approaches, are applied on the fusion task, usually, the vectorial representation of the original information is obtained in such a way the loss function is directly minimized concerning the classification/regression final task. In this case, the compression stage may produce a not self-consistent vectorized representation of the feature space, making it not suitable for modified prediction strategies or multipurpose extensions. On the other hand, as regards ensembling strategies, while combining predictions provided by simple models has been proven to be a robust forecast methodology, ensemble obtained as weighted combinations may underestimate the ability of a specific model to appropriately describe certain dynamics of the target time series rather than others, while multi-input multi-channel networks, usually ensembling neural sub-models, do not benefit from the diversity of models of different nature.

With these premises, this work proposes a two-step forecasting framework based on multi-sequence fusion. As regards the data fusion stage, exogenous time series (in this case *air-quality* and *weather* sequences) space is compressed to a self-consistent representation in order (i) to reduce the noise due to the high-dimensional feature space generated by the multiple time series, (ii) to preserve interconnected time patterns between feature themselves and (iii) to obtain a low-dimensional vector representation suitable for different forecasting strategies. As forecast methodology, an ensemble strategy based on a hybrid neural network exploits features extracted from original data and the predictions provided by different Machine Learning algorithms, used as regressors. This structure allows to connect them to the temporal patterns in the sequence to be predicted, and so to produce robust forecasts taking advantage of different models characteristics. In addition, this modular concept allows to easily extend the ML prediction layer, without modifying neither the compression stage nor the combining one. This characteristic enhances the flexibility of the whole workflow, augmenting the range of contexts it can be applied on.

3. Correlation and causality analysis

The original EHR dataset contains, for each row, a booked appointment of a patient (identified as a unique number corresponding to an encrypted version of its National Identification Number) on a specific day in the regional health department under consideration, with its status (*booked*, *confimed* or *cancelled*). Other information as the physician's encrypted, unique code or the referral identification number has been ignored for our purposes. Time series concerning *respiratory diseases medical services* is then extracted considering the daily number of *confirmed* bookings from 2014 to 2019. Missing data in particular dates are filled with zeros since a missing day on the dataset means no provisions in that specific day.

Since the time series present large weekly variations, the question which arises is whether or not the forecasting can be boosted with external knowledge, as the *air-quality* and the *weather condition* information: validating the hypothesis that relations exist between the *respiratory diseases* sequence and the external time series considered is necessary to design a fusion strategy that preserves such relations, in order to exploit them in the forecasting phase. In particular, we are interested in the following quantities:

- PM2.5 (particulate with diameter less than 2.5 μm, in $\mu g/m^3$);
- PM10 (particulate with diameter less or equal than 10 μm, in $\mu g/m^3$);
- CO (carbon monoxide, in mg/m^3);
- NO2 (nitrogen dioxide, in $\mu g/m^3$)
- mean temperature, in °C;
- total precipitation, in mm;
- wind speed, obtained from the 100 m u-component and v-component of the wind, in m/s.

The *air-quality* data comes from ARPAC-CEMEC (Agenzia Regionale Protezione Ambientale della Campania - CEntro MEteorologico e Climatologico) and include the mean measurements from various stations of Naples. The *weather* measurements are instead taken from the ERA5 hourly data provided by Copernicus Climate Change Service Climate Data Store [61]. All these time series were daily sampled to be coherent with the *respiratory diseases* sequence: in both cases, the daily time series have been obtained as the mean over the 24 h of the considered quantities, with the exceptions of *mm of total precipitation* that was reported as the incremental sum over the 24 h, and so the last value for each day has been taken into account, and of *wind speed*, whose daily mean has been obtained from the hourly module calculated from its *u-component* and *v-component*.

The considered period starts on May 1st, 2017 and ends on April 30th, 2019, obtained as the intersection of the available periods in the provided time series. In Table 1 a summary of the entire dataset of considered sequences for the forecasting problem is reported.

Table 1
Summary of all the time series considered for the prediction of *respiratory diseases bookings* (*RDB*) temporal sequence.

Date	RDB TS	Air-Quality TS				Weather TS		
		$PM2.5$ ($\mu g/m^3$)	$PM10$ ($\mu g/m^3$)	CO (mg/m^3)	NO2 ($\mu g/m^3$)	$\bar{T}$ (C)	TP (mm)	WS (m/s)
2017/05/01	0	7.362	16.441	0.448	27.161	14.699	0.0337	1.532
2017/05/02	251	7.982	22.0477	0.505	27.041	15.133	1.565	1.063
2017/05/03	326	8.664	23.812	0.531	23.668	15.621	0.957	1.118
2017/05/04	331	9.641	26.659	0.523	28.497	15.565	0.0327	1.194
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2019/04/30	153	6.583	12.733	0.446	28.732	12.211	0.0758	1.272

In this Section, the following convention will be used: with X^t we denote a time-dependent random variable in a set $\mathcal{X}$, with $t \in \mathbb{Z}$, while X^{t-l} is the lagged version of X^t with l time steps, with $l > 0$. Finally, we denote with X^t_τ the realization of the random time-dependent variable at step τ.

A first encountered problem is represented by the presence of seasonal high-frequency components: as evidenced from Fig. 1 on the left side, weekly seasonality is strongly present, due to the absence of bookings in the weekend. Regarding the other time series, the *temperature* data presents a yearly seasonality which instead is not evident in the *total precipitation* and the *wind speed*. Considering the *air-quality*, all the time series show a weak weekly seasonality, probably connected to traffic congestion in workdays. Since the relation discovery requires that any form of short seasonality must be removed on all the sequences, a *7-samples* moving average has been applied.

Moreover, since the statistical techniques employ causality tests, which rely on the concept of stationarity, we make sure that all the considered time series are first-order differenced as many times as required, when necessary. In particular, the adopted criterion in checking the stationarity of all the time series is the Augmented Dickey–Fuller test [62]: this test checks the presence of a unit root in the time series, with the null hypothesis representing the presence of a unit root, rejected if the resulting statistics is within a critical value range; in our case, the chosen critical value is in the 95th percentile (p-value ≤ 0.05). After checking the test, if the null hypothesis cannot be rejected, all the time series are first-order differenced, then the test is executed again. For our deseasonalized time series, one differentiation is sufficient. The preprocessing results can be seen in Fig. 1 on the right side.

After making all the time series stationary, tools belonging to two main statistical areas are used for extrapolating the statistical relations: *correlation analysis* and *causality analysis*. The correlation expresses the strength of a relationship if a fixed function (in particular the linear function for the Pearson correlation) is used to estimate a variable in terms of another one. In contrast, causality is used to measure the power as regressive data of a sequence concerning another one's prediction. Given that some of the provided time series may present delayed effects, it is necessary to analyse the time-lagged relations. Considering the correlation analysis, the *time-lagged cross-correlation* provides a first panoramic of the capability in finding synchronicity between the two time series. Here the time lag is taken into account. In time series analysis, such function is defined as $\rho(X^t, Y^t, k) = \rho(X^t, Y^{t-k})$, for $k \in \mathbb{Z}$, where $\rho(X, Y)$ is a chosen correlation measure between two generic time series X and Y.

Regarding the causality analysis, we used the following approaches: *Granger causality*, *Convergent Cross Mapping* (CCM), and *PCMCI*$^+$. These methodologies have been used in various contexts, earth sciences [63] in particular. As evidenced in the literature, the usage of different techniques is needed to understand and suppress possible artefacts caused by each method's drawbacks.

3.1. Granger causality

Granger causality [64] was one of the first attempts to characterize the predicting causality between two phenomena in a statistical way. Originally, such a concept was applied to statistical random processes and not for time series. Indeed, as indicated by Eichler et al. [65], Granger causality is based on the following two fundamental assumptions: (i) the effect does not precede the cause in time; (ii) all the available information is contained in the provided causal series (universality). More formally: X^t *Granger-causes* Y^t if X^{t-l} improves the prediction capability of Y^t for $l = 1, \dots, L$ if a suitable model is employed. The most simple way to test the Granger causality is by estimating the following linear autoregressive model:

$$AR: \quad Y^t = c_1 + \sum_{l=1}^{L} \alpha_l Y^{t-l} + \sum_{l=1}^{L} \beta_l X^{t-l} + R_1^t$$

The null hypothesis to reject is $H_0 : \beta_1 = \beta_2 = \cdots = \beta_L = 0$. In this case, the statistical test is the F-test computed on the residual errors of two autoregressive models: one with all the coefficients, another with the coefficients β_l set to zero.

This method has the advantage of being very simple and is widely used in various science fields, but has many drawbacks: first of all, the test is parametric with the assumption of linearity. Secondly, the assumption of universality is hard to be met. In the case of our study, we applied this methodology to find the relationship between the regressors and the *respiratory diseases booking* sequence. As shown in Fig. 2, the air-quality time series present long (greater than ten days) causality effects, while the weather measures show only very short causality effects.

3.2. Convergent cross mapping

In 2011, Sugihara et al. ideated a new causality technique which relies on dynamical systems terminology. *CCM* works by considering the time series as components of a discrete dynamical system. Differently from Granger causality, it can be used to find causality in coactive systems [66]. The principles behind this approach are that it is possible to extrapolate information through the state space obtained by state reconstruction techniques.

In particular, by Takens theorem, a particular state-space manifold can be reconstructed from a single component of the dynamical system, in our case a univariate time series. If a time series X^t causes Y^t, then the values of X^t can be predicted from the embedded version of Y^t obtained from the reconstruction, and such procedure is called "cross-mapping".

In details, let $\tau < N$, with N the total available time steps, let $\hat{Y}^t$ be the embedded version of Y^t, and let us consider any forecasting algorithm which is trained on a subset of X^t, called the library, containing all the points until time step τ, and let $\rho_{CCM}(Y^t, X^t, \tau)$ be the prediction skill, i.e. the measure of the prediction correlation on the complementary subset of the training set. Then, the causality from X^t to Y^t can be characterized as follows:

- $\rho_{CCM}(Y^t, X^t, \tau)$ is monotonic as τ increases;
- $\rho_{CCM}(Y^t, X^t, \tau)$ saturates to a value $\rho_{CCM}(Y^t, X^t)$ far from 0, as τ is closer to N;
- if the opposite $\rho_{CCM}(X^t, Y^t, \tau)$ is not monotonic or, if it saturates, the limit $\rho_{CCM}(X^t, X^t)$ is significantly less in magnitude than $\rho_{CCM}(Y^t, X^t)$, then the causality is unidirectional.

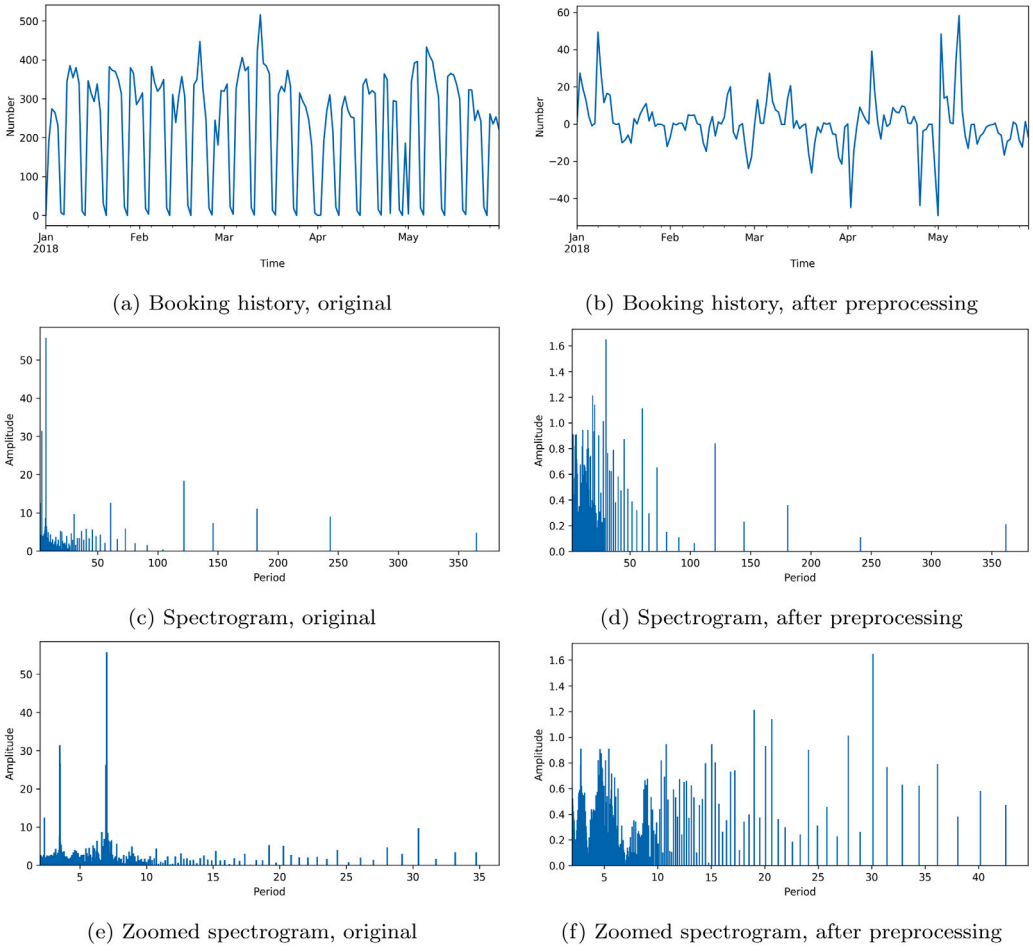


Fig. 1. Time series of *respiratory diseases medical service*. The left side represents the original data, while the right side is obtained after the preprocessing phase. The top 2 plots show the first six months of 2018. The remaining plots represent the spectrograms with the periods on the x axis, instead of the frequencies.

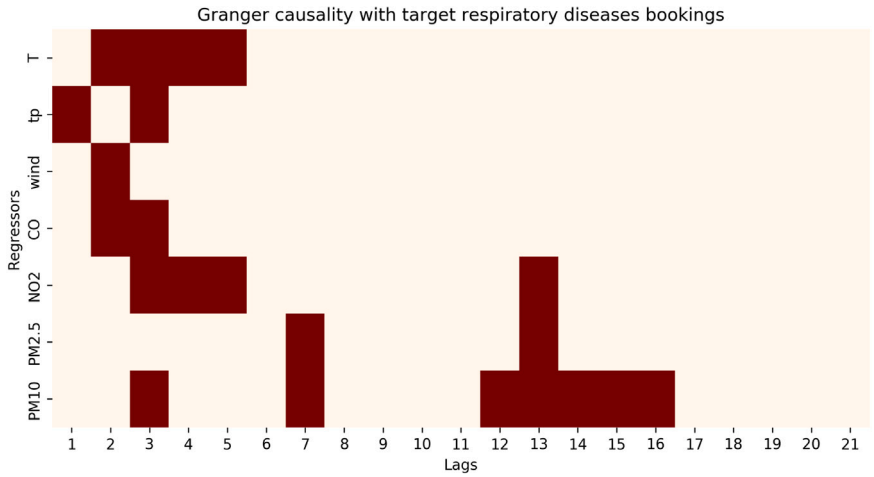


Fig. 2. Heatmap of the Granger causality between the regressors (columns) and the *respiratory diseases bookings* time series at different lags (rows). Dark colour indicates that a regressor time series Granger-causes the bookings, light pink otherwise. The threshold for the positive response is obtained imposing the p-value less than or equal to 0.05.

The results of this technique on the dataset under consideration can be seen from Fig. 3, where the predictive capability of the regressors causing the respiratory diseases converges to a positive value, though weak, while the opposite relations remains closely irrelevant. Regarding the regressors, two kinds of situations are particularly relevant: (i) fairly strong and bidirectional causality, found with the *air-quality* time series and also between *wind speed* and *CO*, with the *wind speed* causing *CO* relation stronger (bottom right); (ii) weak one-directional causality, as seen for *total precipitation* influencing the particulate concentration (bottom left).

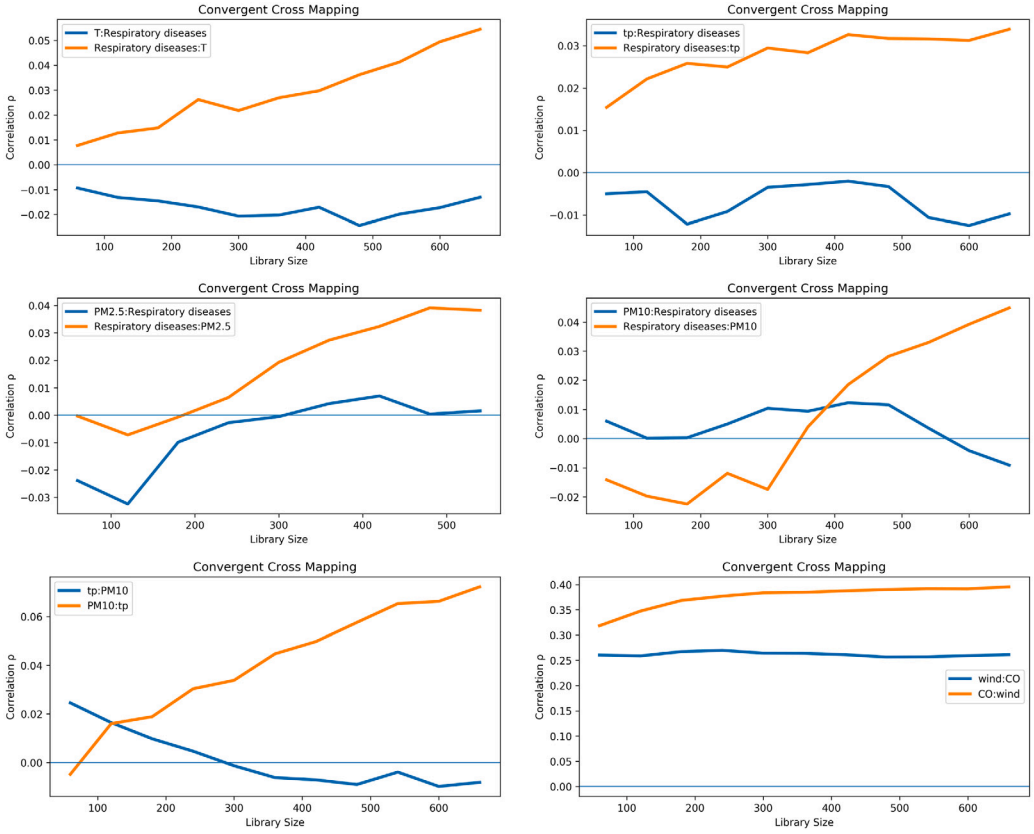


Fig. 3. Results of CCM on the various time series. The first four images represent the relation between *respiratory disease booking* time series and four of the exogenous sequences: *temperature* (top left), *total precipitation* (top right), *PM2.5 concentration* (mid left), *PM10 concentration* (mid right). The last two represent the relations between the regressors: *total precipitation* and *PM10* (bottom left), *wind speed* and *CO* (bottom right), respectively.

3.3. PCMCI$^+$

PCMCI$^+$ [67], which is an improved version of PCMCI [68], is the state of the art of the causal discovery, where all the possible relations between the different time series are considered in the form of a multivariate time-series graph. Here the nodes are the time lags of the temporal sequences, and the links represent the causal connections. This procedure is made of various steps: the first one, called condition selection, picks for each node of the graph a selection of the nodes to be considered as parent nodes through conditional independence tests, then momentary conditional independence (MCI) test finds the effective relations by verifying the conditional independence with the previously selected series. The other steps are needed to find contemporaneous relations. In general, conditional independence tests can be done with various methodologies, but we focused on the Partial correlation test, which evaluates the conditional probabilities relevance with the partial correlation function.

The application of this technique to our study is shown in Fig. 4. Such graphical relation finds weak relations between the regressors (except for *PM2.5-PM10*, positive, and *wind speed - NO2*, negative) and only a few and weak relations from the *air-quality* data to the *respiratory disease bookings* time series, with only the *CO* has a direct causality on the respiratory diseases.

4. Prediction workflow

Analysis carried out in Section 3 shows existing relations between the three groups of time series chosen for the predictive task: in particular, the series of *respiratory diseases bookings* appears to be influenced by both *weather* time series and *air-quality* ones, although in a non-strictly-linear fashion. Moreover, connections exist between *weather* TS and *air-quality* TS themselves; in particular, the second group appear to be influenced by the first one. In this context, the idea is to exploit such *weak short-to-mid-term* relations to provide an enhanced *short-term* prediction of the *respiratory diseases bookings* sequence: due to the complex shape of the relations mentioned above, no straightforward methodology in extracting predictive information on interactions between the considered sequences is possible; therefore, we designed a data-driven framework, based on Supervised Learning methodologies, able to extract such relations and provide robust and reliable *7-days ahead* forecasts through a Hybrid Neural Network, whose aim is to combine predictions obtained through Machine learning regressive algorithms and features extracted from the sequences as discussed in the following. Moreover, since the lagged version of the exogenous time series, used as input of the predictive stack of the pipeline (together with autoregressive and periodicity features of the *respiratory diseases bookings* sequence), generates a high dimensional feature subspace, a LSTM Autoencoder has the aim of reducing the dimension of such feature set, preserving relations between the *weather* and the *air-quality* sequences.

4.1. Preprocessing stage

Guided by the preliminary analysis, in this stage of the workflow, all the regressive variables are created. In the case of the *respiratory diseases bookings* sequence, it can be observed the presence of a strong multimodal periodic pattern, shown by spectrogram analysis, connected to yearly and weekly seasonalities. These time series characteristics are exploited in the regression problem generating categorical variables representing seasonalities of the data. Moreover, features to consider weekends, country and local holidays are also added to the dataset. Besides, we expect that predicted values depend on short-term patterns extracted from lagged variables of the Time Series itself.

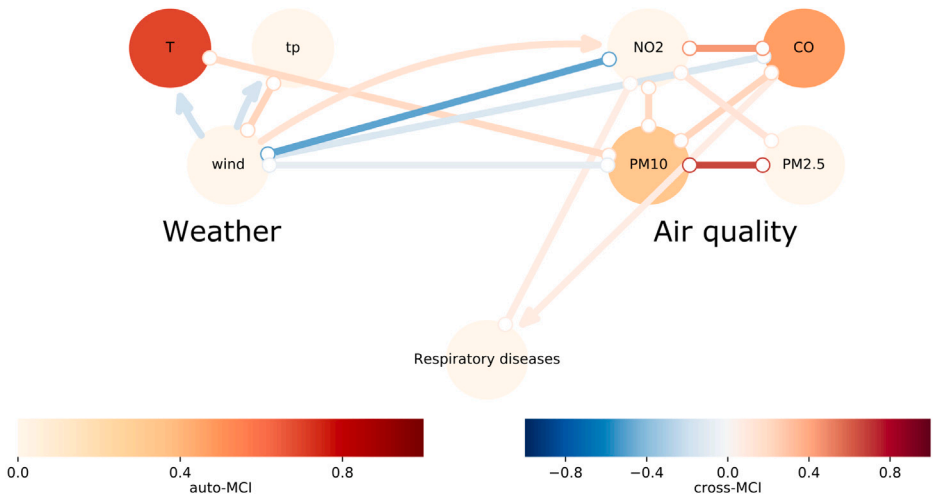


Fig. 4. Graph of causality relations obtained with *PCMCI*$^{+}$. The links with an arrow indicate direct causality towards the pointed variable, while the non-arrowed links indicate unclear direction of causality. The more the link is blue, the more adverse is the effect, while the more the link is red, the causality is proactive. According to this graph, weak relations from the air quality measurements to the *respiratory diseases bookings* are present. At the same time, the weather influences only the air quality (negatively for the *wind speed* (wind), while positive for *temperature*). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

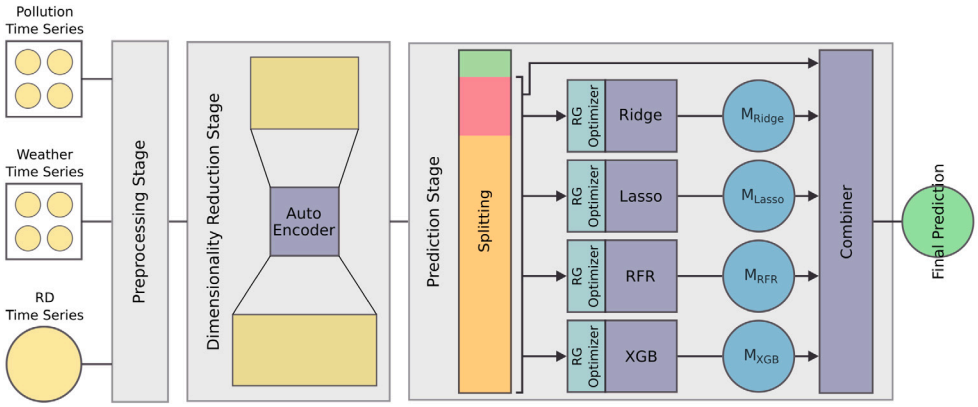


Fig. 5. Schematic design of the proposed framework: temporal sequences of *respiratory diseases bookings*, *weather* and *air-quality* variables are preprocessed as discussed in Section 4.1; then the dimensionality reduction strategy, described in Section 4.2 is applied on the subset composed by lagged variables extracted from exogenous time series; Finally, the obtained dataset is split in *train/validation/test* sets, used to train, optimize and validate the whole procedure: the machine learning predictors are trained over the *train set* (in yellow), their hyperparameters are optimized through a *random grid search*, and their forecast on the validation set (in red) fed the *Combiner*, with lagged and periodicity features, which is trained in recognizing patterns between Time Series attributes and forecasts produced by predictive models. The final forecast is then obtained on the *test set* (in green). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The autocorrelation analysis shows a strong connection between each sample and the signal's behaviour in the preceding seven days. This relation remains remarkable on multiples of this period. To exploit this behaviour for obtaining the *7-step-ahead* forecast, lagged variables obtained from *lag7* to *14* are also considered in the regressive task.

Finally, to take into account the *weak relations* between the three groups of sequences, lagged variables from each of the *weather* and *air-quality* sequences are added to the feature space. Since both linear and non-linear causality tests highlighted a not-homogeneous short-to-mid correlation among the time series of these two sets, *lag0* to *28* from *weather* sequences and *lag7* to *28* from *air-quality* sequences are chosen as regressive variables of the forecasting problem. It is worth underlining that the framework is designed to predict seven days in the future, so *lag0* to *lag6* data of *weather* series are, in principle, unavailable, since they correspond to values not yet collected. However, weather forecasts are proven to be reliable within seven days; hence the experiments presented in Section 5 mimic the situation in which the final forecast is obtained exploiting also available weather forecasts.

In more mathematical terms, all the preprocessing phase can be represented as a mapping of the type:

$$y_t \rightarrow \mathcal{H}(t, \Theta)$$

where y_t express the *target* value of the *respiratory diseases bookings* sequence for the Supervised Learning problem. So, given y time series to be predicted and $\tilde{y}_i$, with $i = i, \dots, k$, exogenous sequences, the map $\mathcal{H}$ associates each y_t to the features $([y_{t-7}, \dots, y_{t-s}], [\tilde{y}_{1,t-s'}, \dots, \tilde{y}_{1,t-28}], \dots, [\tilde{y}_{k,t-s''}, \dots, \tilde{y}_{k,t-28}], \Theta_t)$ where, in our case, $s = 14$, $s' = 0$ in the case of *weather* series, $s'' = 7$ if *air-quality* sequences and Θ_t is a vector expressing periodicities related to the sample t.

4.2. Dimensionality reduction

Preprocessing stage described in Section 4.1 generates a high-dimensional feature space where, as pointed out before, relations also exist between $\tilde{y}_i$ sequences. Moreover, the relations appear to be weak, and might be not recognized by ML algorithms used as first predictors (details about the prediction stage are discussed in Section 4.3) due to the well-known issue of "curse of dimensionality". In order to preserve a *data-driven* approach and to overcome *noise-related* and dimensionality problems, the dataset has been processed through an LSTM-Autoencoder.

Firstly introduced in the 1980s [Rumelhart et al. 1986], Autoencoders are Neural Networks addressing the task of mapping an input

Table 2
Hyperparameter configuration for the AutoEncoder and the Combiner. Regarding the AutoEncoder, n_{regr} represents the total number of features of the dataset generated from exogenous time series.

		Layers	Units	Kernel sizes	Activation	Value
Autoencoder	Encoder LSTM	3	$(n_{regr}, n_{regr}/2, 36)$	–	*tanh*	–
	Decoder LSTM	3	$(36, n_{regr}/2, n_{regr})$	–	*tanh*	–
	Learning rate	–	–	–	–	$1e-4$
	Regularization	–	–	–	–	$l_2(5e-8)$
	Batch size	–	–	–	–	256
Combiner	LSTM	2	(96, 48)	–	*tanh*	–
	Conv1D	4	(8, 8, 8, 8)	(2, 3, 5, 7)	*ReLU*	–
	Dense	3	(64, 32, 1)	–	*ReLU*	–
	Learning rate	–	–	–	–	$1e-3$
	Regularization	–	–	–	–	$l_2(8e-3)$
	Batch size	–	–	–	–	32

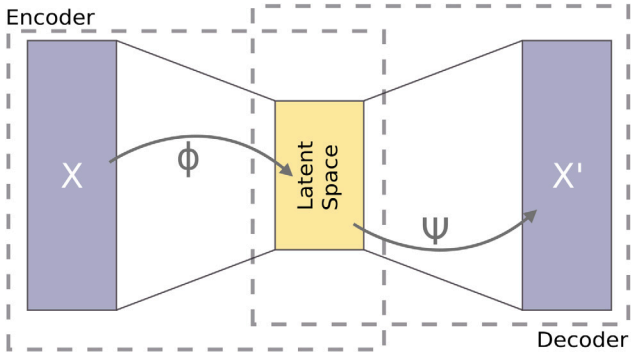


Fig. 6. Simplistic structure of an Autoencoder. The *encoder* maps the original input to its compressed representation, the *latent space*. The *decoder* extracts information from such space to reconstruct the original input, or, at least, an its good approximation. After the training phase, the low-dimensional representation of the feature space is the *latent space* itself.

space X in space X', which has to be at least a good approximation of the first one. It is composed by two main parts: (i) the *encoder*, which maps the input in a lower-dimensional space called *latent space* or *code* that contains a compressed representation of the original data obtained through a bottleneck layer of artificial neurons, (ii) the *decoder*, which extract the only information it is allowed to use to reconstruct the input as more accurate as possible. In mathematical terms, defined $\Phi : X \to \mathcal{L}$ the encoding function which maps the input space X to the *latent space* $\mathcal{L}$ and named $\Psi : \mathcal{L} \to X'$ the decoding function which operates on the latent space and returns values in the output space of the network X', in the case the square loss is chosen as cost function, Φ and Ψ are such that:

$$\Phi, \Psi : \arg\min_{\Phi,\Psi} \|X' - (\Psi \circ \Phi)X\|^2 \tag{1}$$

where $dim(X) = dim(X') > \mathcal{L}$. As can be observed, since the target is automatically defined by the input, it is typically defined as a *self-supervised* one (see Fig. 6).

Proposed by Hochreiter & Schmidhuber [69], LSTM cells were introduced to face the problem of vanishing gradient of standard Recurrent Neural Network. Due to the introduction of gates that control the information to be neglected and to be stored in the state of the cell itself each time new one is provided to the network, these structures are capable of retain knowledge about sequential processes granting the stability of backpropagation mechanism. In particular, the gates that compose a LSTM base unit can be written as:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t]) \tag{2}$$

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t]) \tag{3}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t]) \tag{4}$$

where W weighs the output of the previous LSTM block h_{t-1} at time $t-1$. Subsequently, the equations for cell state, candidate cell state and final output are:

$$\tilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t]) \tag{5}$$

$$C_t = f_t \circ C_{t-1} + i_t \circ \tilde{C}_t \tag{6}$$

$$h_t = o_t \circ \tanh(c_t) \tag{7}$$

where $\tilde{C}_t$ is a candidate for the cell state at time t, C_t is the internal state of the unit and h_t is the final output of the cell. In this form The long term dependencies and relations are encoded in the cell state vectors (see Fig. 7).

The main idea in combining LSTM cells in an Autoencoder structure is to the benefit of memory capacity of this particular type of recurrent neuron in the *latent space* construction. Since the feature sub-space of lagged variables from exogenous data (from *weather* and *air-quality* sets) is built starting from sequences connected by non-linear weak relations, the LSTM-Autoencoder guarantees a dimensionality reduction procedure through a projection in a low-dimensional space in which both temporal auto-relations and connections among different time series are taken into account.

4.3. Predictive stage

After the compression, the last stage of the pipeline reported in Fig. 5 has the aim of producing forecasts exploiting the new dataset, whose generic row related to the sample t can be expressed as:

$$y_t \to \left([y_{t-7}, \dots, y_{t-s}], \Gamma_t, \Theta_t \right) \tag{8}$$

where Γ_t is a vector containing the latent representation of variables, at timestamp t, derived from the aforementioned procedure. As it can be observed, this stage comprises three main operational blocks: (i) the splitting process of the dataset, (ii) a first layer of ML predictive algorithms providing the initial forecasts, and (iii) a Hybrid Neural Network producing the final prediction by combining the previously described data.

4.3.1. Splitting phase

The dataset is split in three sections according the time direction, namely *train set*, *validation set* and *test set*, defined, in relation to the targets, as follows:

$$\mathcal{D}_{train} \to [y_1, \dots y_{N-(N_{val}+N_{test})}]^T$$
$$\mathcal{D}_{val} \to [y_{N-(N_{val}+N_{test})+1}, \dots, y_{N-N_{test}}]^T$$
$$\mathcal{D}_{test} \to [y_{N-N_{test}+1}, \dots, y_N]^T$$

where N_{val} and N_{test} assume the meaning of *validation* and *test* set dimensions, respectively, while N represent the original dimension of the preprocessed dataset.

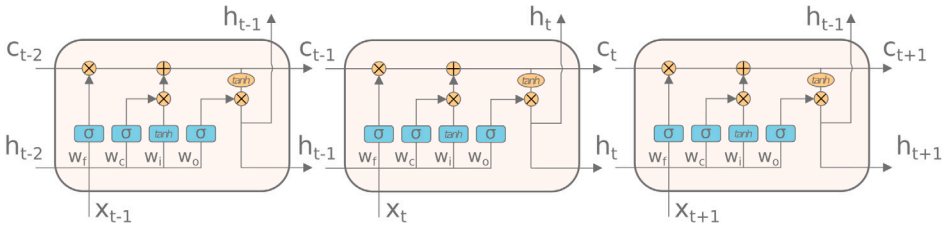


Fig. 7. Schematic representation of LSTM cells. Inside each cell four gates can be found. Their configuration allows a filtering process on information kept and forwarded during the train phase, limiting the problem of vanishing gradient common to standard RNNs.

4.3.2. Machine learning predictors

By taking into account the preliminary analysis results, the selected algorithms are two tree-based regressors. In particular, a Random Forest regressor and a Gradient Boosted Decision Trees Regressor have been chosen. Besides, two linear models, namely a Ridge and a Lasso Linear predictors, are also considered. Tree-based algorithms have been chosen due to their intrinsic ability to model complex patterns and express non-linear relations, inherent characteristics of the problem faced in this work. Also, to preserve a certain degree of variability among the used models, the first two ones have been joined by two linear predictors that, despite their simplistic hypothesis in modelling relations among variables, have been proven to perform unexpectedly well in the forecasting domain.

During the "learning" phase of the workflow these predictors are trained on $\mathcal{D}_{train}$, and they produce forecast on $\mathcal{D}_{val}$. Since the more accurate is the first forecast, the more reliable the final prediction of the *Combiner*, each of ML models is optimized in the sense of Hyperparameter Optimization (HO) concerning the prediction provided on the validation set. In particular, the methodology chosen for HO is a Random Grid Search. Once ML algorithms are optimized, the set $\mathcal{D}_{test}$ is used as feature space to obtain the forecasts that will be merged in the *Combiner*.

4.3.3. Hybrid neural combiner

In the described workflow (see Fig. 5), the last phase of the predictive stage is done by a Hybrid Neural Network. The aim is twofold: a multi-input CNN branch has the task of recognizing short-term behaviours and relations between the forecasts provided by ML algorithms, while a stacked LSTM one detects mid-to-long-term temporal dependencies between the features provided in input and described in Section 4.3. This hybrid architecture is trained on the validation set $\mathcal{D}_{val}$, in such a way, it can learn how to exploit relations between lagged variables, exogenous sequences, periodicity features and preliminary forecasts. Then the final prediction is obtained feeding the *Combiner* with the features, and ML forecasts related to the test set $\mathcal{D}_{test}$.

Convolutional Neural Networks (CNN), introduced by LeCun et al. [70], have the purpose of extracting information about local connections by performing discrete convolution operations, as the name suggests, on the input matrix. These operations, performed on K filters $W \in \mathbb{R}^{L_M \times L_N}$, where L_M and L_N are the dimensions of the convolution kernel, generate weight matrices expressing the more relevant features of the data. In our case the input is shaped as a bi-dimensional tensor $X = (X_t)_{t=0}^{N-1}$, with X_t a vector corresponding to the characteristic features; therefore, the convolution operation, in the case of a single channel, can be written as:

$$s^1(i,h) = \left(w_h^1 * x\right)(i) = \sum_{j=-\infty}^{\infty} w_h^1(j,1)x(i-j,1) \tag{9}$$

where $w_h^1 \in \mathbb{R}^{1\times k\times 1}$ is the filter and $s^1 \in \mathbb{R}^{1\times N-k+1\times M_1}$ is the feature map. This output is then passed through a non-linear function $\mathcal{F}(\cdot)$ to give $y^1 = \mathcal{F}\left(s^1\right)$. In this framework, due to the characteristics of shared weights and operational localization, the number of parameters to learn is considerably reduced and, subsequently, this allows to extrapolate translation-invariant patterns among features.

The dense tail of the proposed hybrid network shown in Fig. 8, is composed by neural structures made by fully connected layers of perceptrons, not linked with each other in the single layer. In general, the output of a dense block composed of N_{in} input neurons and N_{out} output neurons can be written as:

$$y_i = \mathcal{F}\left(\sum_{j=1}^{N_{\text{in}}} w_{ij}x_j + \theta_i\right),\ i = 1,\dots,N_{\text{out}}, \tag{10}$$

where $y \in \mathbb{R}^{N_{\text{out}}}$ is the vector output of the block, x_j are the jth input, w_{ij} is the weight associated to the $(i,j)-th$ link for $i = 1,\dots,N_{\text{out}}$, $j = 1,\dots,N_{\text{in}}$, $\theta_i \in \mathbb{R}$ is the bias, and $\mathcal{F}$ is the non-linear *activation function*, in general a sigmoid function. Thanks to the *universal approximation theorem*, in principle this kind of structure is able to represent a wide variety of functions: in our specific context, it is used to model how to combine information extracted from the two branches to obtain the most reliable overall prediction.

Summarizing, the intuition behind this network architecture is to exploit CNN's local pattern recognition abilities and LSTM memory characteristics (more specifically described in Section 4.2) to take advantage from the predictive capabilities of machine learning algorithms. By using the forecast provided to the net as "corrector" components on the forecast obtained exploiting both endogenous and exogenous features, and combining this information through the final dense block of layers, we expect to enhance the robustness of methodology in providing the final *7-days-ahead* prediction.

5. Results

This section contains a description of the methodologies used for the validation of the proposed methodology, and a discussion on the obtained results. By considering the predictions, we are interested in comparing the forecasts of the hybrid combiner network obtained without exogenous variables and with them, both in the case they are compressed through the *Autoencoder* and considered without compression.

5.1. Experimental setup

Given the temporal sequential nature of the time series, the validation of our predictive methodology has been performed with a modified version of the Nested k-fold Cross-Validation, based on [71] and [72]. As it can be seen in Fig. 9, the size of the validation set and the test set are fixed, while the training set increases at each fold, such that the training set at step $j+1$ includes a number $N_{test} = dim(\mathcal{D}_{test})$ of samples of the validation set used at step j, for $j = 1,\dots,k$. In particular, for our tests, $N_{val} = dim(\mathcal{D}_{val}) = 28$, i.e. four weeks, and $N_{test} = 7$ to provide *7-days* ahead predictions.

Since our framework has been designed to deal with a temporal sequence that always present the same scale, and due to the presence of zero valued samples (linked to weekends and holidays) it was considered appropriate to adopt absolute error measures rather than measures based on the percentage errors [73]. As evaluation metrics, the Mean Absolute Error (MAE), the Root Mean Squared Error (RMSE), the R-squared coefficient (R^2) have been chosen. Both MAE and RMSE measure the error magnitude, while R^2 measures the variability taken into account by the prediction in comparison with the truth.

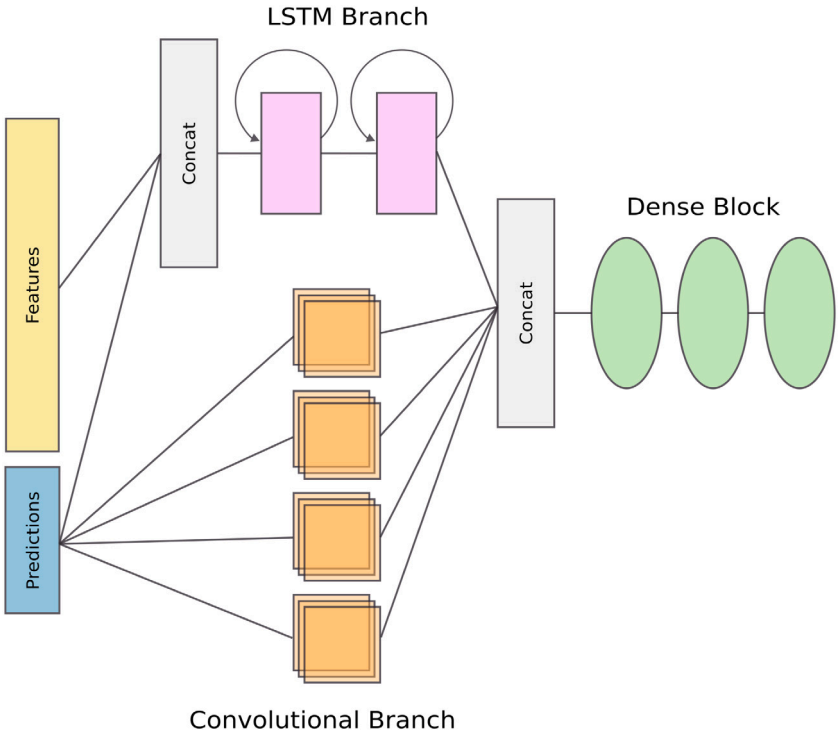


Fig. 8. Sketch of the Hybrid Neural Network used to generate the final prediction. The LSTM branch detects long-term temporal patterns between the input variables while the CNN layers catch the short-term relations between the predictions provided by machine learning algorithms. The two branches' outcomes are joined through a stack of Dense layers that calibrate the importance between these two sets of information and provide the *7-step-ahead* prediction.

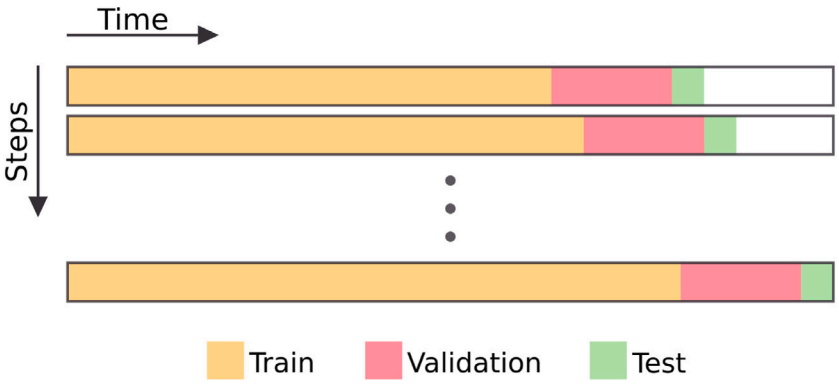


Fig. 9. Procedure of iterative splitting of the time series related dataset for the k-fold cross validation. At step j, with $j = 1, \dots, k$, the size of the training set is $N - (k - j)N_{test}$, with N_{test} the size of the test set, therefore at step $j + 1$ the training set will contain additional N_{test} time steps which were in the test set of step j. The prediction returned by the proposed hybrid neural combiner is evaluated on the test set (in green). In particular, for our tests, we adopted $k = 5$.

Table 3
Comparison of 5-Fold Cross Validation results for *7-days ahead* forecasts between single predictors (Lasso and Ridge, Boosted Trees and Random Forest predictors presented in Section 4.3.2), models and combination strategies reported in Section 5.3, and the proposed framework according to error measures presented in this Section. The measures are expressed as mean ± standard deviation of the k-fold results. Best values are highlighted in bold, while second best ones are underlined.

Predictor	Baseline			All Regressors			Encoded Regressors		
	R^2	MAE	$RMSE$	R^2	MAE	$RMSE$	R^2	MAE	$RMSE$
ARIMA	−0.065 ± 0.825	41.454 ± 16.730	49.369 ± 22.037	–	–	–	–	–	–
SARIMA	0.618 ± 0.397	34.118 ± 16.424	41.014 ± 22.893	–	–	–	–	–	–
ETS	0.607 ± 0.424	34.530 ± 19.776	41.410 ± 24.448	–	–	–	–	–	–
ARIMA-ANN	0.177 ± 0.705	38.580 ± 21.934	46.545 ± 27.007	–	–	–	–	–	–
SARIMA-ANN	0.610 ± 0.415	33.907 ± 16.894	40.644 ± 23.666	–	–	–	–	–	–
ETS-ANN	0.587 ± 0.454	33.910 ± 19.993	40.919 ± 25.008	–	–	–	–	–	–
MLP	0.106 ± 0.568	41.217 ± 20.253	51.677 ± 17.313	0.399 ± 0.176	53.794 ± 8.541	62.286 ± 12.109	0.419 ± 0.177	49.981 ± 10.015	55.632 ± 10.944
LSTM	0.526 ± 0.293	42.309 ± 11.669	49.309 ± 10.131	0.643 ± 0.134	46.401 ± 7.047	55.192 ± 5.515	0.681 ± 0.078	47.264 ± 8.834	51.861 ± 8.933
SVR	0.491 ± 0.217	52.338 ± 14.216	58.707 ± 13.597	0.552 ± 0.098	47.622 ± 8.497	57.459 ± 8.176	0.654 ± 0.151	52.252 ± 10.891	57.595 ± 7.741
Lasso	0.627 ± 0.131	50.731 ± 10.365	59.021 ± 11.042	0.749 ± 0.067	46.360 ± 6.367	52.360 ± 8.360	0.733 ± 0.095	46.071 ± 7.279	54.527 ± 9.417
Ridge	0.648 ± 0.135	48.997 ± 10.302	57.276 ± 10.473	0.749 ± 0.070	49.473 ± 7.797	55.061 ± 9.295	0.747 ± 0.086	46.226 ± 7.688	54.644 ± 10.147
XGB	0.776 ± 0.112	29.034 ± 11.780	39.342 ± 15.659	**0.80 ± 0.087**	25.782 ± 10.206	35.849 ± 11.359	0.786 ± 0.070	26.237 ± 3.338	34.903 ± 4.854
RFR	0.679 ± 0.156	33.509 ± 5.266	44.027 ± 8.795	−0.210 ± 0.841	58.905 ± 9.791	68.195 ± 14.139	0.563 ± 0.276	36.712 ± 11.661	47.881 ± 14.494
Ensemble	Baseline			All Regressors			Encoded Regressors		
	R^2	MAE	$RMSE$	R^2	MAE	$RMSE$	R^2	MAE	$RMSE$
Average	0.717 ± 0.110	38.009 ± 10.300	46.888 ± 12.075	0.764 ± 0.077	36.651 ± 6.814	43.201 ± 8.643	0.759 ± 0.1017	37.608 ± 5.514	44.413 ± 7.498
Median	0.709 ± 0.112	38.501 ± 10.762	47.822 ± 11.330	0.760 ± 0.079	38.983 ± 8.360	46.094 ± 9.084	0.751 ± 0.0896	35.470 ± 6.771	44.358 ± 8.556
PSO	0.541 ± 0.365	24.365 ± 8.340	31.402 ± 10.569	0.248 ± 0.428	31.329 ± 6.713	38.476 ± 7.177	0.637 ± 0.1774	24.225 ± 4.181	30.163 ± 5.484
KF	0.672 ± 0.080	35.914 ± 9.428	43.705 ± 7.268	0.714 ± 0.064	34.340 ± 3.537	41.257 ± 4.441	0.693 ± 0.0505	34.750 ± 1.616	42.127 ± 2.935
GA	0.577 ± 0.306	23.908 ± 6.691	30.714 ± 9.138	0.436 ± 0.341	28.487 ± 5.146	35.646 ± 4.408	0.605 ± 0.2083	24.420 ± 4.139	30.741 ± 5.573
Combiner	**0.817 ± 0.117**	**20.707 ± 5.665**	**26.184 ± 7.175**	0.746 ± 0.209	**23.926 ± 10.418**	**30.177 ± 13.002**	**0.855 ± 0.139**	**18.088 ± 6.082**	**21.544 ± 7.699**

Regarding the setup of the Autoencoder and the Combiner networks, Tables 2 contain the hyperparameters configuration. This is the best choice obtained through extensive heuristic experimentations with different configurations. Moreover, the hyperparameters of the Combiner were not changed in all the experimentation regarding the different setup of the features.

5.2. Encoding

As mentioned in Section 4.2, the compression takes place before the prediction stage: it is executed once, before the k-fold procedure. The results of the encoding can be observed in Fig. 10, which shows the reconstruction error on the decoding of the compressed features. The overall mean error, including the standard deviation, resulted in 0.01218 ± 0.00136. Given that the regressors were rescaled between 0 and 1, the lagged values are reconstructed with good precision. The striped peaks indicate that only some specific values in reconstructing regressors are mispredicted in all the lagged components of the *weather* and *air-quality* sequences.

5.3. Predictions

To assess the forecasting performances of the proposed strategy, its results, in terms of the error metrics presented in Section 5.1 on the 5-fold cross validation, have been compared with other widely used methods, resumed as follows:

- univariate forecast models: classical ARIMA, SARIMA and ETS, and their hybrid variants, ARIMA-ANN [74], SARIMA-ANN [75] and ETS-ANN [76] (with the autoregressive components optimized through an automatic grid search on the hyperparameters based on the Akaike Information Criterion);
- Machine Learning algorithms: Support Vector Regression (SVR), Multi Layer Perceptron (MLP), vanilla LSTM, all optimized in a similar manner of the procedure described in Section 4.3.2, and single predictors used in the framework itself (Lasso, Ridge, Boosted Trees and Random Forest);
- weighting ensemble methods applied to the same predictors as the combiner: average and median [77], Particle Swarm Optimization (PSO) [78], Genetic Algorithm (GA) [79], and Kalman Filter (KF) [80] as time-variant weighing model.

As it can be observed in Table 3, the proposed ensemble strategy always better performs, in terms of average MAE and RMSE, of both single regressors and other combination methodologies, significantly reducing the errors on predicted values. As regards the R^2 measure, a distinction must be drowned: it can be noticed that, in the baseline case, i.e. without considering exogenous sequences, as well as in the case *weather* and *air-quality* series are taken into account as compressed feature subspace, the Hybrid Neural Network is able to recover an improved temporal pattern for the *respiratory diseases bookings* time series, resulting in an improved model as concerns this metric. In the case the dimensionality reduction procedure is neglected, even if the R^2 measure of the prediction provided by our ensemble strategy does not result to be the highest, it is close to the better ones, confirming the capability of the *Combiner* in using the forecast of machine learning models as enhancing variables and recognizing, among them, the most reliable ones.

Moreover, the comparison between the results provided by the three different configuration shows how the models behave concerning the graft of exogenous time series in the regression problem. As can be noticed from the Table, in the case no compression is done, despite the high-dimensional feature space both the linear regression models, the boosted tree regressor produce better results compared to the baseline. At the same time, the deep learning single models (MLP, LSTM) and the SVR do not improve their average error measures, but they are able to better model the time series patterns as can be observed from the improved R^2 measure, which affects the standard deviation of MAE and RMSE that decreases with respect to the baseline case, even if RMSE and MAE results, in average, worse to the former case. The Random Forest, instead, is poorly able to provide a reliable prediction, as can be observed in particular from its mean R^2 score and its extreme variability, evidenced by the very high standard deviation. While these last behaviours are attributable to the "curse of dimensionality" and to the presence of noise in the extended dataset, the good attitude of Lasso, Ridge and XGB has two possible explanations: i) the linear predictors, Lasso and Ridge, are both based on the inclusion of a penalization term in the loss function, namely a Manhattan norm and a squared norm of the weights, so the coefficients which are associated to the irrelevant features are truncated in magnitude; similarly for the XGB algorithm, where the strategy of boosting iteratively select models that ignore noisy feature subspace regions; ii) the predictor's hyperparameter optimization produces different configurations than the baseline case, testifying the adaptability of the tuning to the feature space.

All the ensemble strategies provide similar results to the ones attained in the baseline case, with a generally reduced variability of the error measures due to the good performances of three of the four methods considered for the combination, while the *Combiner* performs worse with respect to the baseline, both in terms of the error measures mean values and in terms of their variability, even if it obtains, in average, the best RMSE and MAE values among compared methods. This difference in behaviours are connected to the methodology of the combination: while a weighted combination only exploits predictions provided by the machine learning algorithms, the *Combiner*, taking into consideration the predictions, the aggregated dataset and the features extracted from the *respiratory diseases booking* series, suffers from the high dimensionality of the feature space, and so the noise contained in the variable subset related to *weather* and *air-quality* lagged variables negatively impacts on the temporal pattern reconstruction.

The last part on the right of the Table shows the error measures in the case the encoding strategy is applied on the lagged variables generated from external time series. As can be noticed, the dimensionality reduction does not affect very much the single regressors as regards the average MAE and RMSE, that remain similar to the one obtained without compression. Moreover, the Random Forest recovers its predictive abilities behaving only slightly worse than the baseline situation. The same holds for the weighting strategies, whose performances are comparable with the ones obtained without compression, only in some cases (PSO and GA) better due to the recovered Random Forest prediction. This confirms that the LSTM-Autoencoder preserves the weak relations found with the analysis in Section 3 in the dimensionality reduction process. In this case, instead, the *Combiner* is capable of increasing its predictive abilities exploiting such relations and enhancing its performance according to all the three error metrics considered. With respect to the baseline case, even though the R^2 score is increased by a minimal margin, the MAE and $RMSE$ scores are significantly lower: our proposal attains to the overall best performance among the compared models, providing the most reliable ans robust prediction also in comparison with other combination strategies, as can be also observed in Fig. 11 with reference to the RMSE.

In Fig. 12 is reported a visual comparison between forecasts obtained through the machine learning regressors, then exploited for the combining stage, and the prediction provided by the *Combiner*. Also, the ground truth is shown as a dotted black line. As can be observed the Hybrid Network presented in this work always produce a reliable *7-days-ahead* prediction, as evidenced by the coherent reproduction of temporal fluctuations (as confirmed by the high R^2 score) of the real data. In all the cases can be noticed how the proposed ensemble strategy is able to recover patterns along time direction, exploiting also the information extracted from the time series itself, even when the ML predictors fail in predicting them (as results evident in particular around zero-valued samples of fold 1, expressing weekends); moreover,

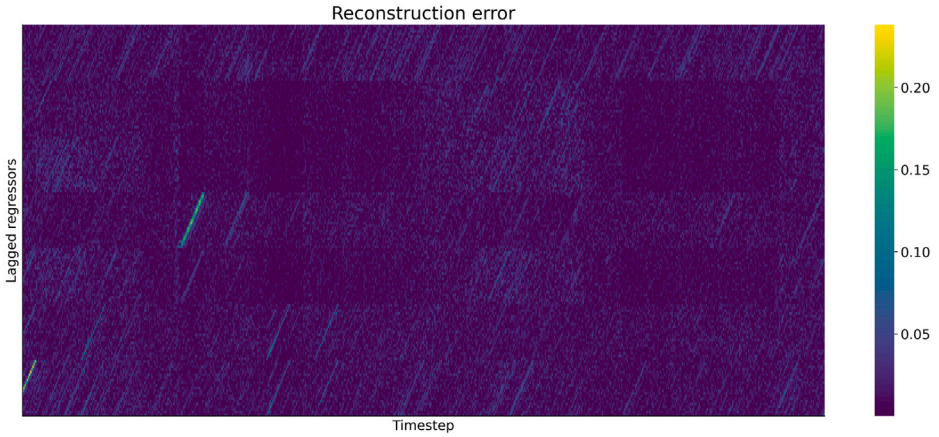


Fig. 10. Heatmap of the reconstruction error, computed as the absolute difference between the original regressors and the reconstructed regressors. Darker colour means lower reconstruction error.

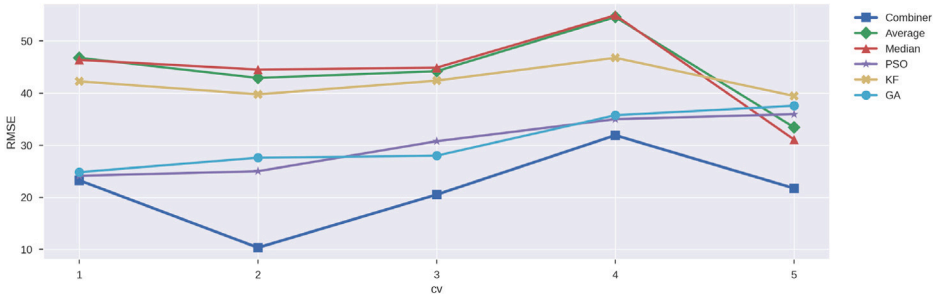
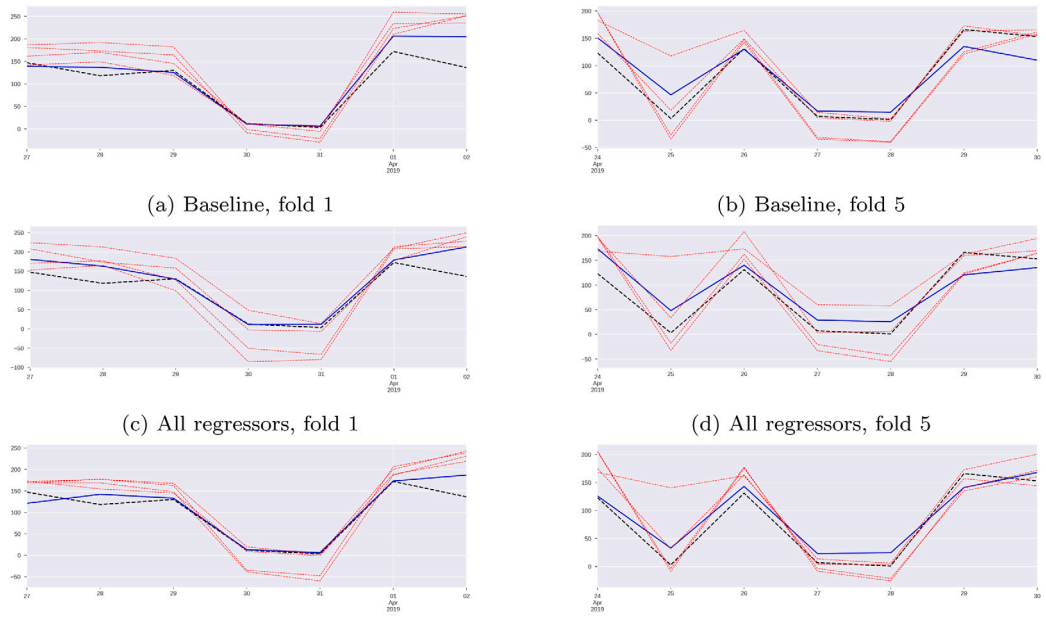


Fig. 11. Comparison of the RMSE for the different combination strategies with respect to each fold of the 5-fold cross validation procedure. Predictions are obtained in the case the low-dimensional feature space, obtained as described in Section 4.2, is used.



(a) Baseline, fold 1 (b) Baseline, fold 5

(c) All regressors, fold 1 (d) All regressors, fold 5

(e) Encoded regressors, fold 1 (f) Encoded regressors, fold 5

Fig. 12. Predictions obtained on the *test set* by the *Combiner* compared with ones provided by the ML algorithm used to feed it. In particular, plots of the forecast for the first and last fold of the cross validation procedure are reported. Dashed black line is the truth, the red lines represent the predictors, the blue line is the *Combiner* forecast.

the augmented predictive power exploiting weak relations between exogenous time series results in an overall enhanced accuracy, which can be highlighted, as an example, in both the first and the last two samples of Fig. 12(f), where a better coherence with respect to the original values is recovered compared to Figg. 12(b), 12(d).

6. Conclusions and future works

In modern organization management, time series prediction has become a key task of decision-making activities. Simulating future situations can make a substantial difference in developing approaches to face them efficiently, even more in the healthcare context where the choice of a strategy rather than another one can produce critical effects. In this context, short-to-mid-term predictions of facilities workloads greatly impact healthcare management decisions, allowing a more adequate and well-balanced organization of structures and workers, optimizing the used resources and providing a better service for patients.

In this context, this paper describes a predictive framework whose aim is to provide robust and reliable forecasts of through deep learning data fusion and ensemble strategies: features extracted from the time series under investigation are joined with the feature space constituted by exogenous temporal sequences through an LSTM-Autoencoder; then, a layer composed of Machine Learning models trained on the joint dataset provides different predictions that are exploited, as regressors, by a hybrid neural network, which combines such forecasts with original time series related data to provide the final *7-days* ahead prediction.

As reported in Section 5, the proposed data fusion strategy provides a low-dimensional feature space which avoids the "curse of dimensionality" issue and reduces the noise without a feature selection procedure. Simultaneously, it preserves the weak relations, identified through a preliminary statistical analysis, between the sequence under investigation and the candidate causal ones. On the other side, the proposed ensemble strategy, exploiting the machine learning forecasts and information extracted from the time sequences, can reliably recover and model temporal patterns, providing a final prediction with enhanced accuracy. As can be observed from the obtained results, all the workflow is proven to outperform widely used forecasting models and combination strategies, attaining significant improvements in short-term predictions of the analysed *respiratory diseases bookings* time series.

Finally, even though the pipeline has been designed to face the specific problem presented in this work, the characteristics of robustness and accuracy exhibited by the framework, together with the flexibility provided by the modular architecture, pave the way to further studies on the proposed approach: the usage of Variational Autoencoders, returning a regularized latent space, can be an option in obtaining an enhanced compression stage, while including new, and more sophisticated, models in the machine learning predictor layer can further improve the accuracy of the final forecast. In this sense, future works will focus on deeply analysing how the framework performs concerning different datasets, comparing it with an extensive collection of state-of-the-art data fusion and forecasting methods, to better assess our approach's effectiveness in the more general context of multi-source time series forecasting.

CRediT authorship contribution statement

Francesco Piccialli: Idea, Methodology, Writing - review & editing, Supervision, Project administration. **Fabio Giampaolo:** Investigation, Data curation, Writing - review & editing. **Edoardo Prezioso:** Conceptualization, Formal analysis, Writing - review & editing. **David Camacho:** Review & editing. **Giovanni Acampora:** Review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the *CUP in un click* - CUP-in-One, research project [Regione Campania - Bando RIS3 2018 - Fase 2 - CUP: B63D18000550007]. The authors would like to thank the M.O.D.A.L. research laboratory (http://www.labdma.unina.it/index.php/modal/) for the effort and support.

References

[1] J. Chen, K. Li, K. Bilal, A.A. Metwally, K. Li, P. Yu, Parallel protein community detection in large-scale PPI networks based on multi-source learning, IEEE/ACM Trans. Comput. Biol. Bioinform. (2018) 1, http://dx.doi.org/10.1109/TCBB.2018.2868088.

[2] G. Xiao, K. Li, Y. Chen, W. He, A.Y. Zomaya, T. Li, CASpMV: A customized and accelerative SpMV framework for the sunway taihulight, IEEE Trans. Parallel Distrib. Syst. 32 (1) (2021) 131–146, http://dx.doi.org/10.1109/TPDS.2019.2907537.

[3] Y. Chen, K. Li, W. Yang, G. Xiao, X. Xie, T. Li, Performance-aware model for sparse matrix-matrix multiplication on the sunway taihulight supercomputer, IEEE Trans. Parallel Distrib. Syst. 30 (4) (2019) 923–938, http://dx.doi.org/10.1109/TPDS.2018.2871189.

[4] C. Liu, K. Li, M. Song, J. Zhao, K. Li, T. Li, Z. Zeng, Coexe: An efficient co-execution architecture for real-time neural network services, in: 2020 57th ACM/IEEE Design Automation Conference, DAC, 2020, pp. 1–6, http://dx.doi.org/10.1109/DAC18072.2020.9218740.

[5] F. Piccialli, V.D. Somma, F. Giampaolo, S. Cuomo, G. Fortino, A survey on deep learning in medicine: Why, how and when? Inf. Fusion 66 (2021) 111–137, https://doi.org/10.1016/j.inffus.2020.09.006.

[6] N. Mehta, A. Pandit, S. Shukla, Transforming healthcare with big data analytics and artificial intelligence: A systematic mapping study, J. Biomed. Inform. 100 (2019) 103311, https://doi.org/10.1016/j.jbi.2019.103311.

[7] M. Malik, S. Abdallah, M. Ala'raj, Data mining and predictive analytics applications for the delivery of healthcare services: a systematic literature review, Ann. Oper. Res. 270 (1–2) (2018) 287–312, http://dx.doi.org/10.1007/s10479-016-2393-z.

[8] S.L. Harris, J.H. May, L.G. Vargas, Predictive analytics model for healthcare planning and scheduling, European J. Oper. Res. 253 (1) (2016) 121–131, https://doi.org/10.1016/j.ejor.2016.02.017.

[9] A. Nelson, D. Herron, G. Rees, P. Nachev, Predicting scheduled hospital attendance with artificial intelligence, NPJ Digit. Med. 2 (1) (2019) 26, http://dx.doi.org/10.1038/s41746-019-0103-3.

[10] A. Alahmar, E. Mohammed, R. Benlamri, Application of data mining techniques to predict the length of stay of hospitalized patients with diabetes, in: 2018 4th International Conference on Big Data Innovations and Applications, Innovate-Data, 2018, pp. 38–43.

[11] H.J. Jiang, C.A. Russo, M.L. Barrett, Nationwide Frequency and Costs of Potentially Preventable Hospitalizations, 2006: Statistical Brief #72, Agency for Healthcare Research and Quality (US), Rockville (MD), 2006.

[12] C.C. Holt, Forecasting seasonals and trends by exponentially weighted moving averages, Int. J. Forecast. 20 (1) (2004) 5–10, http://dx.doi.org/10.1016/j.ijforecast.2003.09.015.

[13] G.E.P. Box, G. Jenkins, Time Series Analysis, Forecasting and Control, Holden-Day, Inc., USA, 1979.

[14] G.P. Zhang, V. Berardi, et al., Time series forecasting with neural network ensembles: an application for exchange rate prediction, J. Oper. Res. Soc. 52 (6) (2001) 652–664.

[15] M.M. Noel, B.J. Pandian, Control of a nonlinear liquid level system using a new artificial neural network based reinforcement learning approach, Appl. Soft Comput. 23 (2014) 444–451, https://doi.org/10.1016/j.asoc.2014.06.037.

[16] Y. Chen, B. Yang, J. Dong, Time-series prediction using a local linear wavelet neural network, Neurocomputing 69 (4) (2006) 449–465, https://doi.org/10.1016/j.neucom.2005.02.006.

[17] A. Jain, A.M. Kumar, Hybrid neural network models for hydrologic time series forecasting, Appl. Soft Comput. 7 (2) (2007) 585–592, https://doi.org/10.1016/j.asoc.2006.03.002.

[18] C.H. Aladag, E. Egrioglu, C. Kadilar, Forecasting nonlinear time series with a hybrid methodology, Appl. Math. Lett. 22 (9) (2009) 1467–1470, https://doi.org/10.1016/j.aml.2009.02.006.

[19] H. Liu, C. Chen, H. qi Tian, Y. fei Li, A hybrid model for wind speed prediction using empirical mode decomposition and artificial neural networks, Renew. Energy 48 (2012) 545–556, https://doi.org/10.1016/j.renene.2012.06.012.

[20] Z. Shen, Y. Zhang, J. Lu, J. Xu, G. Xiao, A novel time series forecasting model with deep learning, Neurocomputing 396 (2020) 302–313, https://doi.org/10.1016/j.neucom.2018.12.084.

[21] J. Chen, K. Li, H. Rong, K. Bilal, K. Li, P.S. Yu, A periodicity-based parallel time series prediction algorithm in cloud computing environments, Inform. Sci. 496 (2019) 506–537, https://doi.org/10.1016/j.ins.2018.06.045.

[22] A. Galicia, R. Talavera-Llames, A. Troncoso, I. Koprinska, F. Martínez-Álvarez, Multi-step forecasting for big data time series based on ensemble learning, Knowl.-Based Syst. 163 (2019) 830–841, https://doi.org/10.1016/j.knosys.2018.10.009.

[23] D. Shaub, Fast and accurate yearly time series forecasting with forecast combinations, Int. J. Forecast. 36 (1) (2020) 116–120, M4 Competition, https://doi.org/10.1016/j.ijforecast.2019.03.032.

[24] C. Gómez-Quiles, G. Asencio-Cortés, A. Gastalver-Rubio, F. Martínez-Álvarez, A. Troncoso, J. Manresa, J.C. Riquelme, J.M. Riquelme-Santos, A novel ensemble method for electric vehicle power consumption forecasting: Application to the Spanish system, IEEE Access 7 (2019) 120840–120856.

[25] P. Musikawan, K. Sunat, Y. Kongsorot, Wind power forecasting using a heterogeneous ensemble of decomposition-based nnrw techniques, ECTI Trans. Comput. Inf. Technol. (ECTI-CIT) 14 (2) (2020) 122–138, https://doi.org/10.37936/ecti-cit.2020142.239860.

[26] M.H.D.M. Ribeiro, L. dos Santos Coelho, Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series, Appl. Soft Comput. 86 (2020) 105837, https://doi.org/10.1016/j.asoc.2019.105837.

[27] M. Moreira, P. Balestrassi, A. Paiva, P. Ribeiro, B. Bonatto, Design of experiments using artificial neural network ensemble for photovoltaic generation forecasting, Renew. Sustain. Energy Rev. 135 (2021) 110450, https://doi.org/10.1016/j.rser.2020.110450.

[28] B. Alhnaity, M. Abbod, A new hybrid financial time series prediction model, Eng. Appl. Artif. Intell. 95 (2020) 103873, https://doi.org/10.1016/j.engappai.2020.103873.

[29] A. Ranjbar, C. Cherubini, Development of a robust ensemble meta-model for prediction of salinity time series under uncertainty (case study: Talar aquifer), Heliyon 6 (12) (2020) e05758, https://doi.org/10.1016/j.heliyon.2020.e05758.

[30] M. Pulido, P. Melin, Comparison of genetic algorithm and particle swarm optimization of ensemble neural networks for complex time series prediction, in: P. Melin, O. Castillo, J. Kacprzyk (Eds.), Recent Advances of Hybrid Intelligent Systems Based on Soft Computing, Springer International Publishing, Cham, 2021, pp. 51–77, https://doi.org/10.1007/978-3-030-58728-4.

[31] S. Kaushik, A. Choudhury, N. Dasgupta, S. Natarajan, L.A. Pickett, V. Dutt, Ensemble of multi-headed machine learning architectures for time-series forecasting of healthcare expenditures, in: P. Johri, J.K. Verma, S. Paul (Eds.), Applications of Machine Learning, Springer Singapore, Singapore, 2020, pp. 199–216, http://dx.doi.org/10.1007/978-981-15-3357-0_14.

[32] S. El-Sappagh, T. Abuhmed, S. Riazul Islam, K.S. Kwak, Multimodal multitask deep learning model for Alzheimer's disease progression detection based on time series data, Neurocomputing 412 (2020) 197–215, https://doi.org/10.1016/j.neucom.2020.05.087.

[33] L. Yuan, J. Ma, J. Gu, H. Wen, Z. Jin, Featuring periodic correlations via dual granularity inputs structured RNNs ensemble load forecaster, Int. Trans. Electr. Energy Syst. 30 (11) (2020) e12571, https://doi.org/10.1002/2050-7038.12571.

[34] J.S. Armstrong, Combining forecasts, in: Principles of Forecasting, Springer, 2001, pp. 417–439.

[35] S. Makridakis, E. Spiliotis, V. Assimakopoulos, The M4 competition: Results, findings, conclusion and way forward, Int. J. Forecast. 34 (4) (2018) 802–808, https://doi.org/10.1016/j.ijforecast.2018.06.001.

[36] R.J. Linn, D.L. Hall, J. Llinas, Survey of multisensor data fusion systems, in: V. Libby (Ed.), Data Structures and Target Classification, vol. 1470, International Society for Optics and Photonics, SPIE, 1991, pp. 13–29, http://dx.doi.org/10.1117/12.44836.

[37] G. Kelly, Data fusion: from primary metrology to process measurement, in: IMTC/99. Proceedings of the 16th IEEE Instrumentation and Measurement Technology Conference, Cat. No.99CH36309, vol. 3, 1999, pp. 1325–1329.

[38] J. Liu, J. Huang, R. Sun, H. Yu, R. Xiao, Data fusion for multi-source sensors using GA-PSO-BP neural network, IEEE Trans. Intell. Transp. Syst. (2020) 1–16.

[39] Q. Feng, D. Zhu, J. Yang, B. Li, Multisource hyperspectral and lidar data fusion for urban land-use mapping based on a modified two-branch convolutional neural network, ISPRS Int. J. Geo-Inf. 8 (2019) 28, http://dx.doi.org/10.3390/ijgi8010028.

[40] Q. Xiao, X. Bai, P. Gao, Y. He, Application of convolutional neural network-based feature extraction and data fusion for geographical origin identification of radix astragali by visible/short-wave near-infrared and near infrared hyperspectral imaging, Sensors 20 (17) (2020) 4940, https://doi.org/10.3390/s20174940.

[41] K. Lee, S. Cheon, C.O. Kim, A convolutional neural network for fault classification and diagnosis in semiconductor manufacturing processes, IEEE Trans. Semicond. Manuf. 30 (2) (2017) 135–142, http://dx.doi.org/10.1109/TSM.2017.2676245.

[42] Y. Liang, Z. Gao, J. Gao, R. Wang, H. Zhao, Data fusion combined with echo state network for multivariate time series prediction in complex electromechanical system, Comput. Appl. Math. 37 (5) (2018) 5920–5934, https://doi.org/10.1007/s40314-018-0669-4.

[43] Y. Kim, P. Wang, Y. Zhu, L. Mihaylova, A capsule network for traffic speed prediction in complex road networks, in: 2018 Sensor Data Fusion: Trends, Solutions, Applications, SDF, 2018, pp. 1–6, http://dx.doi.org/10.1109/SDF.2018.8547068.

[44] F. Rodrigues, I. Markou, F.C. Pereira, Combining time-series and textual data for taxi demand prediction in event areas: A deep learning approach, Inf. Fusion 49 (2019) 120–129, https://doi.org/10.1016/j.inffus.2018.07.007.

[45] M.Z. Uddin, M.M. Hassan, A. Alsanad, C. Savaglio, A body sensor data fusion and deep recurrent neural network-based behavior recognition approach for robust healthcare, Inf. Fusion 55 (2020) 105–115, https://doi.org/10.1016/j.inffus.2019.08.004.

[46] T. Li, S. Fong, K.K. Wong, Y. Wu, X. she Yang, X. Li, Fusing wearable and remote sensing data streams by fast incremental learning with swarm decision table for human activity recognition, Inf. Fusion 60 (2020) 41–64, https://doi.org/10.1016/j.inffus.2020.02.001.

[47] C. Habib, A. Makhoul, R. Darazi, R. Couturier, Health risk assessment and decision-making for patient monitoring and decision-support using wireless body sensor networks, Inf. Fusion 47 (2019) 10–22, https://doi.org/10.1016/j.inffus.2018.06.008.

[48] W. Qi, A. Aliverti, A multimodal wearable system for continuous and real-time breathing pattern monitoring during daily activity, IEEE J. Biomed. Health Inf. 24 (8) (2020) 2199–2207, http://dx.doi.org/10.1109/JBHI.2019.2963048.

[49] J. Jijesh, et al., A supervised learning based decision support system for multi-sensor healthcare data from wireless body sensor networks, Wirel. Pers. Commun. (2020) 1–19, https://doi.org/10.1007/s11277-020-07762-9.

[50] D. Zhang, C. Yin, J. Zeng, X. Yuan, P. Zhang, Combining structured and unstructured data for predictive models: a deep learning approach, BMC Med. Inform. Decis. Mak. 20 (1) (2020) 1–11, https://doi.org/10.1186/s12911-020-01297-6.

[51] F. Ali, S. El-Sappagh, S.R. Islam, D. Kwak, A. Ali, M. Imran, K.-S. Kwak, A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion, Inf. Fusion 63 (2020) 208–222, https://doi.org/10.1016/j.inffus.2020.06.008.

[52] Y. Zheng, X. Hu, Healthcare predictive analytics for disease progression: a longitudinal data fusion approach, J. Intell. Inf. Syst. 55 (2020) 351–369, https://doi.org/10.1007/s10844-020-00606-9.

[53] H. Li, Y. Fan, Early prediction of Alzheimer's Disease Dementia Based On Baseline hippocampal MRI and 1-year follow-up cognitive measures using deep recurrent neural networks, in: 2019 IEEE 16th International Symposium on Biomedical Imaging, ISBI 2019, 2019, pp. 368–371, http://dx.doi.org/10.1109/ISBI.2019.8759397.

[54] Y. Yoo, L.W. Tang, T. Brosch, D.K.B. Li, L. Metz, A. Traboulsee, R. Tam, Deep learning of brain lesion patterns for predicting future disease activity in patients with early symptoms of multiple sclerosis, in: G. Carneiro, D. Mateus, L. Peter, A. Bradley, J.M.R.S. Tavares, V. Belagiannis, J.P. Papa, J.C. Nascimento, M. Loog, Z. Lu, J.S. Cardoso, J. Cornebise (Eds.), Deep Learning and Data Labeling for Medical Applications, Springer International Publishing, Cham, 2016, pp. 86–94.

[55] S. Purwar, R.K. Tripathi, R. Ranjan, R. Saxena, Detection of microcytic hypochromia using cbc and blood film features extracted from convolution neural network by different classifiers, Multimedia Tools Appl. 79 (7) (2020) 4573–4595, https://doi.org/10.1007/s11042-019-07927-0.

[56] B. Majeed, J. Peng, A. Li, Y. Lin, R.I. Delgado, Forecasting the demand of mobile clinic services at vulnerable communities based on integrated multi-source data, IISE Trans. Healthc. Syst. Eng. 0 (0) (2020) 1–15, http://dx.doi.org/10.1080/24725579.2020.1859305.

[57] P. Rangarajan, S.K. Mody, M. Marathe, Forecasting dengue and influenza incidences using a sparse representation of google trends, electronic health records, and time series data, PLoS Comput. Biol. 15 (11) (2019) e1007518, https://doi.org/10.1371/journal.pcbi.1007518.

[58] Z. Ertem, D. Raymond, L.A. Meyers, Optimal multi-source forecasting of seasonal influenza, PLoS Comput. Biol. 14 (9) (2018) e1006236, https://doi.org/10.1371/journal.pcbi.1006236.

[59] S. Pei, J. Shaman, Aggregating forecasts of multiple respiratory pathogens supports more accurate forecasting of influenza-like illness, PLoS Comput. Biol. 16 (10) (2020) e1008301, https://doi.org/10.1371/journal.pcbi.1008301.

[60] T. Rekatsinas, S. Ghosh, S.R. Mekaru, E.O. Nsoesie, J.S. Brownstein, L. Getoor, N. Ramakrishnan, Sourceseer: Forecasting rare disease outbreaks using multiple data sources, in: Proceedings of the 2015 SIAM International Conference on Data Mining, SIAM, 2015, pp. 379–387, https://doi.org/10.1137/1.9781611974010.43.

[61] ERA5 hourly data on single levels from 1979 to present, in: Copernicus Climate Change Service Climate Data Store (CDS), 2018, http://dx.doi.org/10.24381/cds.adbb2d47.

[62] D.A. Dickey, W.A. Fuller, Distribution of the estimators for autoregressive time series with a unit root, J. Amer. Statist. Assoc. 74 (366a) (1979) 427–431, http://dx.doi.org/10.1080/01621459.1979.10482531.

[63] J. Runge, S. Bathiany, E. Bollt, G. Camps-Valls, D. Coumou, E. Deyle, C. Glymour, M. Kretschmer, M. Mahecha, J. Muñoz, E. Nes, J. Peters, R. Quax, M. Reichstein, M. Scheffer, B. Schölkopf, P. Spirtes, G. Sugihara, J. Sun, J. Zscheischler, Inferring causation from time series in earth system sciences, Nature Commun. 10 (2019) http://dx.doi.org/10.1038/s41467-019-10105-3.

[64] C.W.J. Granger, Investigating causal relations by econometric models and cross-spectral methods, Econometrica 37 (3) (1969) 424–438.

[65] M. Eichler, Causal inference in time series analysis, in: Causality, John Wiley & Sons, Ltd, 2012, pp. 327–354 (Chapter 22), http://dx.doi.org/10.1002/9781119945710.ch22.

[66] G. Sugihara, R. May, H. Ye, C.-h. Hsieh, E. Deyle, M. Fogarty, S. Munch, Detecting causality in complex ecosystems, Science 338 (6106) (2012) 496–500, http://dx.doi.org/10.1126/science.1227079.

[67] J. Runge, Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets, in: J. Peters, D. Sontag (Eds.), in: Proceedings of Machine Learning Research, vol. 124, PMLR, Virtual, 2020, pp. 1388–1397.

[68] J. Runge, P. Nowack, M. Kretschmer, S. Flaxman, D. Sejdinovic, Detecting and quantifying causal associations in large nonlinear time series datasets, Sci. Adv. 5 (11) (2019) http://dx.doi.org/10.1126/sciadv.aau4996.

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

[70] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (11) (1998) 2278–2324, http://dx.doi.org/10.1109/5.726791.

[71] S. Varma, R. Simon, Bias in error estimation when using cross-validation for model selection, BMC Bioinformatics 7 (1) (2006) 91, http://dx.doi.org/10.1186/1471-2105-7-91.

[72] C. Bergmeir, J.M. Benítez, On the use of cross-validation for time series predictor evaluation, in: Data Mining for Software Trustworthiness, vol. 191, Inform. Sci. (2012) 192–213, http://dx.doi.org/10.1016/j.ins.2011.12.028.

[73] M. Shcherbakov, A. Brebels, N. Shcherbakova, A. Tyukov, T. Janovsky, V. Kamaev, A survey of forecast error measures, World Appl. Sci. J. 24 (2013) 171–176.

[74] G. Zhang, Time series forecasting using a hybrid ARIMA and neural network model, Neurocomputing 50 (2003) 159–175, https://doi.org/10.1016/S0925-2312(01)00702-0.

[75] F.-M. Tseng, H.-C. Yu, G.-H. Tzeng, Combining neural network model with seasonal time series ARIMA model, Technol. Forecast. Soc. Change 69 (1) (2002) 71–87, http://dx.doi.org/10.1016/S0040-1625(00)00113-X.

[76] S. Panigrahi, H. Behera, A hybrid ETS–ANN model for time series forecasting, Eng. Appl. Artif. Intell. 66 (2017) 49–59, https://doi.org/10.1016/j.engappai.2017.07.007.

[77] R. Polikar, Ensemble based systems in decision making, IEEE Circuits Syst. Mag. 6 (3) (2006) 21–45, http://dx.doi.org/10.1109/MCAS.2006.1688199.

[78] J. Kennedy, R. Eberhart, Particle swarm optimization, in: Proceedings of ICNN'95 - International Conference on Neural Networks, vol. 4, 1995, pp. 1942–1948, http://dx.doi.org/10.1109/ICNN.1995.488968.

[79] D. Li, L. Luo, W. Zhang, F. Liu, F. Luo, A genetic algorithm-based weighted ensemble method for predicting transposon-derived piRNAs, BMC Bioinformatics 17 (1) (2016) 1–11.

[80] P. Baraldi, F. Mangili, E. Zio, A Kalman filter-based ensemble approach with application to turbine creep prognostics, IEEE Trans. Reliab. 61 (4) (2012) 966–977, http://dx.doi.org/10.1109/TR.2012.2221037.