Prediction of Emergency Department Visits Applying an One Health Approach: Further Investigations

Ismaela Avellino
R&D Researcher
GPI SpA
Trento, Italy
email: ismaela.avellino@gpi.it

Isabella Della Torre
R&D Researcher
GPI SpA
Trento, Italy
email: isabella.dellatorre@gpi.it

Francesca Marinaro
R&D Researcher
GPI SpA
Trento, Italy
email: francesca.marinaro@gpi.it

Andrea Buccoliero
R&D Project Manager
GPI SpA
Trento, Italy
email: andrea.buccoliero@gpi.it

Antonio Colangelo
R&D Director
GPI SpA
Trento, Italy
email: antonio.colangelo@gpi.it

***Abstract*—Proper management of emergency rooms is needed to improve healthcare and patient satisfaction, guiding resource allocation. Predicting access and hospitalisation rates through Machine Learning appears feasible and promising, especially when coupled with air pollution and weather data. This work further investigates, in a more detailed way, a previously presented approach that applied predictive algorithms to data related to Brescia's clinical and environmental data from 2018 to 2022 to predict daily accesses or daily hospitalisations for cardiovascular or respiratory disorders. Starting from the previous work, that analysis was improved and widened to a greater geographical area. The applied algorithms' performances satisfactorily adhere to the actual data, especially when using the Support Vector Machine and Random Forest's models as regressors on daily accesses and respiratory disease-caused hospitalisations. Even if the specific value is not always correctly predicted, generally, the overall trend seems to be rightly forecasted, and performance metrics are rather satisfying. Although additional work could still be encouraged to improve the models' performances, results are rewarding and represent a new point of view on a complex and relevant matter. The real-life application of this One Health approach is now possible and could quite easily be adapted to other areas, too, with the final objective of improving the quality of healthcare and people's quality of life.**

***Keywords-Forecasting; ER accesses; Hospitalisations; Pollution; Weather; One Health; Environmental exposure*.**

I. INTRODUCTION

This work is an extension of our previous research presented at the AIHealth 2024 conference that took place in Athens, Greece [1].

It aimed to enable the forecast of the Emergency Department (ED) and Emergency Room (ER) fluxes of patients based on their geographically fixed short-term exposure to pollution agents and weather conditions.

Here, this approach is further investigated and broadened to a larger geographical area, extending the applied methods and reaching a more detailed analysis.

Properly managing ED and ER is crucial to providing functional healthcare and improving patients' satisfaction [2]. It leads to a strong need for accurate prediction of visitor volume and patient admissions to facilitate the planning of resources and staff for the whole hospital.

Multiple researchers have tried to predict access and admission rates based on historical ED data by creating scores or using Deep Learning (DL) or Machine Learning (ML) models (like Recurrent Neural Networks, Logistic Regression, Random Forest or Extreme Gradient Boosting) to forecast daily accesses to the ER [3]–[5], the possibility of a patient's hospital admission after going through the triage [6] or even the risk of death [7]. Results are so encouraging that others continue to look for associations with the surrounding environment.

There is proof that weather affects one's health, especially for people with specific illnesses or healthcare needs. For example, there seems to be a link between the daily temperature and ED admissions for cardiovascular diseases or significant exacerbation of asthma in adults that visit ED [8][9]. Generally speaking, regarding cardiovascular disorders, a worsening of the patient's well-being and cardiac arrests appear to be influenced by temperature and other stressors like humidity and atmospheric pressure [10][11]. Moreover, there is also proof of links between air pollution and specific illnesses. Substances like $PM_{2.5}$, PM_{10}, NO_x, O_3 and SO_2 influence cardiac arrests [12], cardiac arrhythmia [13], cognitive decline in adult population [14], COVID-19 incidence [15], development of chronic kidney disease [16] or Type 2 diabetes [17]. $PM_{2.5}$ and PM_{10} are also linked to hospital admissions for cardiovascular [18] and respiratory diseases [19]. $PM_{2.5}$ levels also seem to be directly associated with increased daily ED visits for ulcerative colitis [20], while solar radiation is inversely associated with inflammatory bowel disease admissions [21]. There also seems to be a correlation between the number of hospitalised asthma patients and both weather (i.e., temperature and humidity) and pollution (i.e., $PM_{2.5}$, PM_{10} and NO_x) [22]. Finally, ML models (i.e., AutoRegressive Integrated Moving Average and Multilayer Perceptron) have also been used to try to predict accesses to the ER by patients affected by infecting respiratory diseases after being exposed to $PM_{2.5}$ [23].

Some of these investigations are based on long-term exposure to pollution (even 20-years long [14]), while others on a few days or even same day's exposure [15] [18] and some even on both [13].

The amount of days linked to long- or short-term exposure differs for each study and group of researchers, leading to different temporal definitions and freedom of choice when fixing it. For example, when considering only climatic variables, greater exposure can be seven days long [5], meaning that the forecast based on today's data will be projected one week in the future.

Based on these literary pieces of evidence, trying to predict all accesses to the ED or hospitalisation post-triage for specific illnesses, working on climate, pollution and historical accesses time-series belonging to the same area, seems feasible, even if complex.

Indeed, one of the underlying issues of ED visits' prediction is how non-homogeneous and inconstant patients' emergency accesses are. An urgent crisis can suddenly arise without any clear previous sign or from a multitude of variables that are difficult to constantly monitor simultaneously: inpatients' fluxes in ERs and hospitals are ever-changing and subject to the influence of factors like seasons, outbreaks and social conditions [5].

Each year, between 77000 and 80000 patients visit the ER of the largest Brescia hospital [24], and 24% of them get admitted. This is the reason why this ED seemed like the perfect place where to start our attempt at accurately predicting future accesses based on historical and local meteorological and pollution data.

This paper contains a description of the analysed materials and applied methods (i.e., the datasets and the ML approaches applied to them) in Section II, the reached results in Section III, a comment on them in Section IV and a few final remarks in Section V.

II. Materials and Methods

This section describes the study design, analysed datasets (both clinical and environmental data) and applied algorithms.

A. Study Design

This study primarily aims to daily predict the volume of patients going through the ER of a precise hospital in Brescia, Italy.

Forecasting algorithms were designed for ER accesses and hospitalisations from triage for cardiovascular or respiratory diseases.

This retrospective study applies to daily data (clinical and environmental) for a period going from January 1, 2018, to December 31, 2022. A four-year (i.e., 2018–2021) dataset was used to train the forecasting models, while the remaining data were used to test its forecasting capability. The final datasets used to feed the predictive algorithms combine the clinical and the environmental data.

Data collection

The following subsections describe the datasets of interest analysed in this study.

B. Clinical Data

The original clinical dataset was given by a hospital in Brescia to GPI for research purposes.

The dataset contained all anonymous ER access data for the period going from 2018 to 2022. For each access (i.e., a person on a specific day), there were as many rows as the exams the person had undergone; pre-processing was made to have only one row for each ED visit while maintaining the patient's data (like the date of ER visit, their age, sex and zip code of their home address, the list of medical exams they underwent and, in case they went through hospitalisation, their diagnosis as an ICD9-CM code).

The patients came from different cities: most came from the area surrounding the hospital, while others came from other Italian regions or even from abroad. This study's focus was the area for which environmental data had been collected: Brescia. This work presents two different population divisions based on how the Brescia area is geographically identified by the Italian bureaucracy and due to differences in how environmental data were computed to get the best granularity possible. This will be further described in Subsection II-C.

Table I describes the original overall dataset.

TABLE I
Brief description of clinical data.

Year	*Total accesses*	*Median age*	*Male percentage*	*Female percentage*
2018	60176	55	49%	51%
2019	60106	56	49%	51%
2020	47205	58	52%	48%
2021	49571	57	50%	50%
2022	56631	56	51%	49%

In 2018, 12% of patients were below 18 years old, 31% between 19 and 49, 23% between 50 and 69, 34% above 70. In 2019, 11% of patients were below 18 years old, 30% between 19 and 49, 24% between 50 and 69, 35% above 70. In 2020, 9% of patients were below 18 years old, 29% between 19 and 49, 27% between 50 and 69, 35% above 70.
In 2021, 10% of patients were below 18 years old, 30% between 19 and 49, 26% between 50 and 69, 34% above 70. In 2022, 12% of patients were below 18 years old, 29% between 19 and 49, 25% between 50 and 69, 34% above 70. Amongst the most recurrent diagnoses of the hospitalised patients, through all years, were pneumonia and chronic heart failure.

Table II reports the different percentages of ER accesses in the quarters of each analysed year.

The variables included in our final dataset are:

- Categorical information about the date (as described in Table III), from which dummies were computed

TABLE II
DISTRIBUTION OF ER ACCESSES IN THE DIFFERENT YEARS QUARTERS.

Year	***1st quarter***	***2nd quarter***	***3rd quarter***	***4th quarter***
2018	25.7%	25.3%	24.2%	24.8%
2019	26.7%	24.5%	23.7%	25.1%
2020	29.7%	23.4%	24.7%	22.2%
2021	22.7%	24.8%	25.8%	26.7%
2022	23.1%	25.7%	25.0%	26.2%

- Daily number of accesses to the ER or hospitalisations coming from it, limited to those patients coming either from just the city of Brescia or also from its entire province
- The rolling mean of the number of accesses or hospitalisations, applying a seven-day window for calculation.

TABLE III
DESCRIPTION OF CALENDRICAL INFORMATION.

Calendrical variable	***Definition***
Day of the week	Monday, Tuesday, [...], Saturday, Sunday
Day of the month	1, 2, 3, 4, [...], 28, 29, 30, 31
Month	January, February, [...], November, December
Year	2018, 2019, 2020, 2021, 2022

The subdivisions in different pathological groups were done by selecting the correct hospitalisations through the ICD9-CM codes reported as the primary diagnosis for their access.

Table IV describes the dataset restricted to the city of Brescia.

TABLE IV
BRIEF DESCRIPTION OF CLINICAL DATA (CITY OF BRESCIA).

Year	***Total accesses***	***Median age***	***Male percentage***	***Female percentage***
2018	10389	56	46%	54%
2019	10963	58	47%	53%
2020	9835	61	50%	50%
2021	11082	60	49%	51%
2022	12597	60	49%	51%

In 2018, 11% of patients were below 18 years old, 30% between 19 and 49, 24% between 50 and 69, 35% above 70.
In 2019, 10% of patients were below 18 years old, 29% between 19 and 49, 24% between 50 and 69, 37% above 70.
In 2020, 8% of patients were below 18 years old, 27% between 19 and 49, 27% between 50 and 69, 38% above 70.
In 2021, 9% of patients were below 18 years old, 28% between 19 and 49, 25% between 50 and 69, 38% above 70.
In 2022, 11% of patients were below 18 years old, 27% between 19 and 49, 23% between 50 and 69, 39% above 70.

Table V describes the dataset widened to Brescia's province.

TABLE V
BRIEF DESCRIPTION OF CLINICAL DATA (BRESCIA'S PROVINCE).

Year	***Total accesses***	***Median age***	***Male percentage***	***Female percentage***
2018	53378	55	48%	52%
2019	53678	57	49%	51%
2020	43445	59	49%	51%
2021	46386	58	50%	50%
2022	52518	57	50%	50%

In 2018, 12% of patients were below 18 years old, 31% between 19 and 49, 24% between 50 and 69, 33% above 70.
In 2019, 11% of patients were below 18 years old, 30% between 19 and 49, 24% between 50 and 69, 35% above 70.
In 2020, 9% of patients were below 18 years old, 28% between 19 and 49, 27% between 50 and 69, 36% above 70.
In 2021, 10% of patients were below 18 years old, 29% between 19 and 49, 26% between 50 and 69, 35% above 70.
In 2022, 12% of patients were below 18 years old, 28% between 19 and 49, 25% between 50 and 69, 35% above 70.

A little contextualisation of the clinical dataset: it is fundamental to note that the area around Brescia suffered substantially from the outbreak of the COVID-19 pandemic, and the number of cases affected by coronavirus pneumonia far exceeds the occurrences of any other diagnosis during 2020.

It is possible to observe from Table I and Table V, and this is something already reported in other studies [25] [26], that the number of accesses to ER significantly decreased from 2019 to 2020: this is explainable because Italy was in a strict lockdown for several months that year. Hence, it was less likely, for example, for car accidents to happen or for people wearing masks to get the flu.

Note that, regarding our data of interest, while this trend is observable for both the general accesses and those from Brescia's province, it is not valid for the patients from the city itself.

C. Environmental Data

The environmental data have been supplied by the startup Hypermeteo [27] under GPI's specific request to match the spatio-temporal dimension of the already-at-disposal clinical dataset.

The environmental data for the city of Brescia are defined per day and zip (the Italian CAP) code, guaranteeing spatial-temporal precision. On the contrary, the province area is defined by ISTAT codes, while the corresponding zip code was also reported, aggregating them.

Two different codes describe Italian municipalities: CAP and ISTAT. The first is a postal code, while the other links to the homonymous Italian statistics authority [28].

These environmental data were obtained employing a mathematical model with a resolution of $10km$x$10km$, corrected through normalisation and down-scaling, and applied to data by Lombardia's Regional Environmental Protection Agency (ARPA [29]) weather stations.

While the model was built for the entire Lombardia region, environmental data were extracted for the province of Brescia only, and our study was divided into two phases.

In fact, at first, only data from the city of Brescia itself were analysed, and the results of this approach were reported in our previous publication [1]. Now, we have re-approached the same city data but also widened our analysis to the entire province.

When working with the sole city of Brescia, its 15 zip codes were differentiated both in the clinical and environmental data and were all linked to only 1 ISTAT code. This, unfortunately, was not the case for the province data.

In this sense, the city of Brescia and its province are differently identified. While every municipality is linked to one and only one ISTAT code, Brescia's city is further defined into 15 different CAP codes, where one CAP code can define multiple of its province's municipalities.

Since the ISTAT code can be linked to many different CAP codes and the environmental province data was defined based on the former, if we wanted to link the clinical dataset to the environmental one, we had to find a way to reduce the latter to one row per zip code.

For this reason, for the same zip code, we computed the mean of each environmental variable for each day, enabling the later merge between this dataset and the clinical one.

Apart from the different identifying geographical codes, the rest of the datasets are precisely the same for both approaches, describing the same variables.

Specifically, the reported environmental variables are:

- Temperature (min and max values) (T_{min}, T_{max} [$°C$])
- Humidity (min and max percentage values) (RH_{min}, RH_{max} [$\%$])
- Precipitations (Prec [mm])
- PM_{10} and $PM_{2.5}$ [$\mu g/m^3$]
- NO_x, SO_2, NMVOC and O_3 [$\mu g/m^3$]
- Total solar irradiance (SSW_{tot}) [Wh/m^2].

For each variable, safety ranges, provided along with the dataset, were considered in order to give a label (i.e., zero or one) to each value to indicate if a value could be safe or not, respectively. Depending on the variable, either lower or upper bounds were considered, as reported in Table VI. These safety ranges have been chosen with Hypermeteo based on institutional guidelines [30].

TABLE VI
SAFETY RANGES FOR ENVIRONMENTAL VARIABLES.

Environmental variable	Lower and Upper Bounds	
	Min value	*Max value*
NO_x	-	25 $\mu g/m^3$
$PM_{2.5}$	-	15 $\mu g/m^3$
PM_{10}	-	45 $\mu g/m^3$
SO_2	-	40 $\mu g/m^3$
NMVOC	-	1000 $\mu g/m^3$
O_3	-	100 $\mu g/m^3$
T_{min}	-10 $°C$	-
T_{max}	-	35 $°C$
RH_{min}	15%	-
RH_{max}	-	95%
Prec	-	10 mm
SSW_{tot}	-	8500 Wh/m^2

Subsequently, we computed the number of occurrences in which the data were out of range for the city and province datasets. Occurrences are a single day of the five years considered per single zip code.

In the city of Brescia, in around the 71% of occurrences NO_x was out of range, it was the 60% of cases for $PM_{2.5}$, 20% for PM_{10}, 17% for RH_{max}, 8% for the precipitations, 7.5% for O_3, 2% for T_{max}, 0.5% for SSW_{tot}, 0.3% for RH_{min} and 0 cases out of range for NMVOC, SO_2 and T_{min}.

In its province, in around the 44% of occurrences NO_x was out of range, it was the 46% of cases for $PM_{2.5}$, 13% for PM_{10}, 21% for RH_{max}, 8% for the precipitations, 5% for O_3, 0.8% for T_{max}, 0.4% for SSW_{tot}, 0.4% for RH_{min}, 0.4 for T_{min} and 0 cases out of range for NMVOC and SO_2.

The issue of having multiple rows of data for the same date (i.e., one row for each zip code) has been handled similarly as in a project [31] found during our bibliographic research: each environmental variable has been labelled with the zip code it is referred to, and it is used as a column with daily values, thus grouping all data belonging to the same date on one row.

Again, a clarification on the context: the area surrounding Brescia is densely inhabited and industrialised, resulting in one of the most polluted areas in Europe [32].

PREDICTIVE ALGORITHMS

The following subsections describe the different predictive algorithms applied to the analysed datasets: Random Forest, Artificial Neural Network, Support Vector Machine and AutoRegressive Integrated Moving Average.

D. Random Forest

In order to predict future ER accesses or hospitalisations, based on our clinical and environmental data, the first algorithm to be applied was the same as the one used in the previous study [1].

A Random Forest (RF) approach was implemented in Python with the application of the open-source library Scikit-learn [33]. This model was chosen based on an article [34] that applied it to a temperature prediction problem: the analogy with our dataset highlighted this approach as a fascinating candidate for this type of analysis.

RFs apply sequential splits to the data such that the separation is maximised in regards to a homogeneity criterion, resulting in a combination of tree predictors so that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest [4].

The random forest algorithm picks N random records from the dataset and builds a decision tree based on them, repeating the process for the selected number of trees. The topic has been tackled as a regression problem as we have considered the target variable (i.e., daily accesses) as continuous.

The main goal of our modelling approach was to create an algorithm that improved the error compared with the average baseline one (Average Baseline Error, ABE), which we computed as the mean absolute difference between the actual values and the rolling mean. We considered the latter the most basic prediction to be produced since it simply uses the rolling mean of the target variable calculated for the previous seven days as the predicted future value.

To find the best parameters to set the RF model to, we applied a Python optimisation library called Optuna [35] that, through multiple trials, finds the values that minimise or maximise a specific metric of interest. In our case, we opted to minimise the MAE.

Trying to improve performances (both in terms of metrics and computational time), we applied Optuna to obtain the optimal parameters (n_estimators, max_depth, min_samples_split, min_samples_leaf) for the RF. For each case study, we ran multiple trials to reach their best combination.

The results that are reported in Subsection III-A are based on different combinations of the datasets, as we applied the same model on both the city of Brescia's and the entirety of its province's whole datasets, and then, again, for both of them, on different data combinations and only on cardiovascular or respiratory disorders data.

Since we also worked on the entire province, we widened the application of the same logic used in our previous study [1] to its data.

To predict the future values of interest, we, again, applied a temporal lag to the datasets, but, this time, only one day (and not five too). This means that the observed data from the previous day is used to predict the volume of patient accesses or hospitalisations on the subsequent one.

The different analyses that were carried out, trying to improve the model's performance and potentially spot specific influential variables, can be divided into seven macro cases:

1) To enable further discussion over the preciseness of our daily accesses' predictions and validate our datasets' composition, we decided to deepen our analysis on what we considered to be our baseline.
   Thus, in addition to computing the ABE, we also decided to apply a model constructed using the same number of trees as applied during the previous study [1] to each spatial dataset (city and province) reduced to only contain the rolling mean and calendrical information, thus without any environmental feature.
2) The RF algorithm was applied to the initial preprocessed accesses' dataset made up of patients from the city of Brescia:
   a) at first, the applied model was created using the same number of trees as applied during the previous study [1],
   b) then, the best model was searched, and the best combination of its parameters was found in order to produce the best achievable prediction,
   c) finally, this last model was applied to only the two most important (as computed by the best model itself) features.
3) The RF algorithm was applied to the initial preprocessed hospitalisations dataset made up of patients from the city of Brescia and whose main diagnosis was a cardiovascular disease:
   a) at first, the best model was searched and found by optimising its parameters,
   b) then, it was applied to only the two most important (as computed by the best model itself) features.
4) The RF algorithm was applied to the initial preprocessed hospitalisations dataset made up of patients from the city of Brescia and whose main diagnosis was a respiratory disease:
   a) at first, the best model was searched and found by optimising its parameters,
   b) then, it was applied to only the two most important (as computed by the best model itself) features.
5) The RF algorithm was applied to the initial preprocessed accesses' dataset made up of patients from the entire province of Brescia:
   a) at first, the applied model was created using the same number of trees as applied during the previous study [1],
   b) then, the best model was searched, and the best combination of its parameters was found in order to produce the best achievable prediction,
   c) later, trying to improve the metrics, we casually divided the first four years (2018-2021) into train (80%) and test (20%) that we input into a trial for the best model and then used it to predict our last available year (2022, our usual year of test). We have done so as it looked like using a casual division gave better metrics' values,
   d) then, a model with the same configuration as the best one was applied to only the two most important (as computed by the best model itself) features,
   e) finally, the study on the most important features was reapplied, not to create a new RF model but rather to study which environmental features have the most influence on the prediction when discarding the rolling mean or both the rolling mean and calendrical information about the different days.
6) The RF algorithm was applied to the initial preprocessed hospitalisations' dataset made up of patients from the entire province of Brescia and whose main diagnosis was a cardiovascular disease:
   a) at first, the best model was searched and found by optimising its parameters,
   b) then, again, a study on which environmental features have the most influence on the prediction (thus probably also on the hospitalisations) was conducted.
7) The RF algorithm was applied to the initial preprocessed hospitalisations' dataset made up of patients from the entire province of Brescia and whose main diagnosis was a respiratory disease:
   a) at first, the best model was searched and found by optimising its parameters,

b) then, again, a study on which environmental features have the most influence on the prediction (thus probably also on the hospitalisations) was conducted.

To evaluate the performances of our models, we applied different metrics.

First, we computed the ABE to be considered as the value to be improved.

Then, we also computed the mean and standard deviation (that will be reported as dispersion in Section III) of both the actual and predicted values.

Here, the equations for the other metrics applied to evaluate the models' performances are reported. They were Mean Absolute Error (MAE, 1), Mean Absolute Percentage Error (MAPE, 2), Symmetric Mean Absolute Percentage Error (SMAPE, 3), and R² score (4) [36]:

$$MAE = \sum_{i=1}^{N} |\hat{y}_i - y_i| \tag{1}$$

$$MAPE = \frac{100}{N}\sum_{i=1}^{N} \frac{y_i - \hat{y}_i}{y_i} \tag{2}$$

$$SMAPE = \frac{100}{N}\sum_{i=1}^{N} \frac{|\hat{y}_i - y_i|}{(|y_i| + |\hat{y}_i|)/2} \tag{3}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{4}$$

where N is the test sample size, y is the actual values' vector, $\hat{y_i}$ is the predicted values' one and $\bar{y}$ is the mean of the actual test values.

Since the applied algorithm is a regressor, the usual Accuracy equation cannot be used. It was replaced with a value we called Acc* that we computed using the MAPE subtracted from 100% as reported in Equation 5. MAPE is sometimes called Forecast Error Percentage, so it seemed fitting to create such a metric.

$$Acc^* = 100 - MAPE \tag{5}$$

We also plotted the comparison graphs between the actual and predicted values for the daily accesses and hospitalisations. We also plotted their smoothed version to appreciate the preciseness of the forecast more, as data were noisy. The applied smoothing filter was Savitzky-Golay, with a temporal window length of 31 days and a polynomial order of 2.

E. Artificial Neural Network

As in the previous study [1], trying to improve the results given by the algorithm described in Subsection II-D, other algorithms were applied to hospitalisation data.

Specifically, the first was an Artificial Neural Network (ANN) [37] designed in Python. This model was only used on hospitalisation data linked to cardiovascular or respiratory diseases of patients from both the city and the province of Brescia.

Since ANN is a distance-dependent model, trying to achieve the best performance possible, we applied scaling on the data through a specific library [38].

The used model was a 2-layer shallow neural network, and an optimisation algorithm was, again, applied to search for the best parameters possible.

The selected metrics to evaluate the performances were MAE (1) and SMAPE (3).

Once more, we plotted the comparison graphs between the actual and predicted values for the daily cardiovascular hospitalisations and their smoothed version computed by applying the same filter described in Subsection II-D.

F. Support Vector Machine

Further trying to improve the prediction of hospitalisations, a Support Vector Machine (SVM) [39] was implemented.

It is a supervised ML algorithm that, in this case, we used for regression and applied in Python through its homonymous library [40]. Its main aim is to find the optimal hyperplane in an N-dimensional space that can separate the data points in different classes, guaranteeing a margin between the closest points of different classes to be the maximum possible.

When a Support Vector is applied to solve regression problems, its produced model depends only on a subset of the training data because the cost function ignores samples whose prediction is close to their target.

Our implementation applied a Linear Support Vector Regressor fine-tuned through the Python application of a Scikit-learn library called GridSearchCV [41].

This model was applied only to the hospitalisation data for both spatial (city and province) datasets.

The applied metrics to evaluate the performances were MAE (1) and R² score (4).

We plotted the comparison graphs between the actual and predicted values for the daily hospitalisations and their smoothed version computed by applying the same filter described in Subsection II-D.

G. AutoRegressive Integrated Moving Average

In the previous study [1], trying to improve the results given by the algorithm described in Subsection II-D, a specific ML model for multivariate time-series prediction was applied to the hospitalisation data: an AutoRegressive Integrated Moving Average (ARIMA) model [42].

It is a popular algorithm used in time series analysis and forecast.

For the previous analysis, we applied the auto-ARIMA process [43] in Python, while, this time, we applied another library that enabled the guided research of the best parameters: statsmodels' ARIMA function [44].

The basic idea of the ARIMA model is to use a particular mathematical algorithm to describe the random time series of the data and then predict the future values based on the past and present values through a so-called autoregression. An ARIMA (p, d, q) model can be described through Equation 6:

$$(1-\sum_{i=1}^{p}\varphi_i L^i)(1-L)^d X_t = (1+\sum_{i=1}^{q}\theta_i L^i)\varepsilon_t \qquad (6)$$

where L represents the lag operator, p represents the number of autoregressive terms, q represents the number of moving average terms, d represents the degree of differencing, and ϕ, θ and ϵ are relevant parameters.

Since the achieved results were, again, not promising, we are not going to report all of them, but just an interesting aspect about the city's respiratory hospitalisations' time-series prediction from this model that highlights a peculiar characteristic of the actual data.

The reported results come from the application of the ARIMA model on hospitalisations due to respiratory diseases of patients from the city of Brescia.

III. Results

This section reports the obtained results from the various predictive algorithms.

Even if the algorithms have been fed with different datasets, they always include only data related to patients whose home address' zip code is either inside the city of Brescia or its entire province, based on their objective as described in Subsections II-D, II-E, II-F and II-G.

As already declared, the presented results are performance metrics' values or plots.

The second ones show the curves representing the daily predicted values (always plotted in magenta) versus the actual values for the testing year (i.e., 2022), plotted in different colours based on the predictive algorithm they come from.

Note that when metrics could not be computed due to data sparsity, they were not reported for that specific case study.

A. Random Forest

This subsection presents the results of the RF application to our datasets of interest.

City of Brescia

Please note that the results reported for the datasets constituted by accesses and hospitalisations of patients from the zip codes of Brescia (the same dataset analysed in the previous study [1]) have been improved and newly computed.

1) Daily accesses' baseline: As previously anticipated, to further evaluate the goodness of our RF models for the daily accesses' predictions, we analysed datasets reduced to only the rolling mean and calendrical information.

The first metric to be computed was the ABE, as it was considered to be the value to improve. It was equal to 4.92.

Here are the results for the 1000 trees model:

- MAE = 4.83
- R² score = 0.21
- Acc* = 85.63%
- MAPE = 14.37%
- SMAPE = 6.9%
- Mean of actual accesses = 34.51
- Dispersion of actual accesses = 6.69.

2) Daily accesses: The following values are the metrics computed for the model created using the original number of trees (i.e., 1000):

- MAE = 4.75
- MAPE = 14.37%
- SMAPE = 6.9%
- Acc* = 85.63%
- R² score = 0.21
- Mean of predicted accesses = 34.20
- Dispersion of predicted accesses = 3.22.

The following reported values are the metrics computed for the best model coming from the optimisation (i.e., 423 trees):

- MAE = 4.63
- MAPE = 13.91%
- SMAPE = 6.8%
- Acc* = 86.09%
- R² Score = 0.24
- Mean of predicted accesses = 33.89
- Dispersion of predicted accesses = 2.78.

Figures 1 and 2 plot the actual (in blue) and predicted (in magenta) accesses and their smoothed version, respectively.

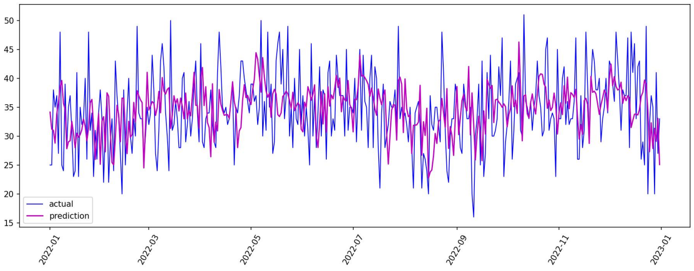

Figure 1. Random Forest's predicted (as computed by the best model) and actual values of daily ER accesses for Brescia.

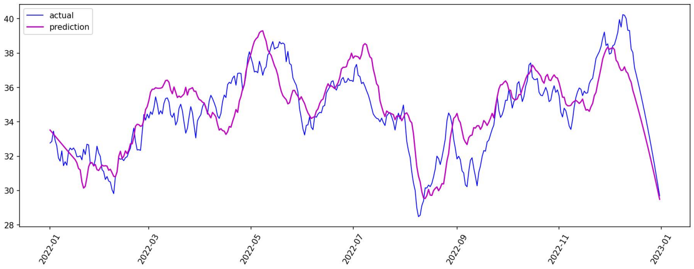

Figure 2. Random Forest's smoothed predicted (as computed by the best model) and actual values of daily ER accesses for Brescia.

The following metrics are the ones computed from the model whose input were just the two most important features (resulting from the best model):

- MAE = 5.56
- Acc* = 82.82%.

These features were the rolling mean and the day of the month.

3) Daily hospitalisations for cardiovascular diseases: We computed the ABE to use as the value to be improved, and it was equal to 0.49.

The following reported values are the metrics computed for the best model coming from the optimisation (i.e., 855 trees):

- MAE = 0.51
- SMAPE = 77.86%
- R² Score: 0.08
- Mean of actual hospitalisations: 0.45
- Mean of predicted hospitalisations = 0.48
- Dispersion of actual hospitalisations = 0.68
- Dispersion of predicted hospitalisations = 0.31.

Figures 3 and 4 plot the actual (in blue) and predicted (in magenta) cardiovascular hospitalisations and their smoothed version, respectively.

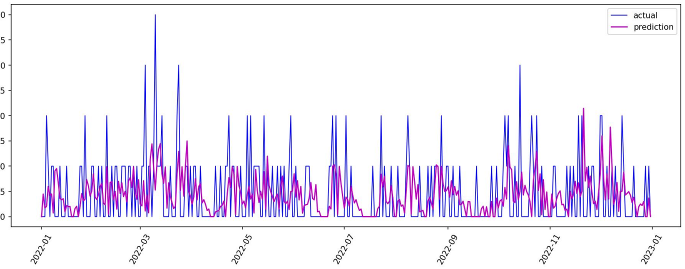

Figure 3. Random Forest's predicted (as computed by the best model) and actual values of daily hospitalisations for cardiovascular diseases for Brescia.

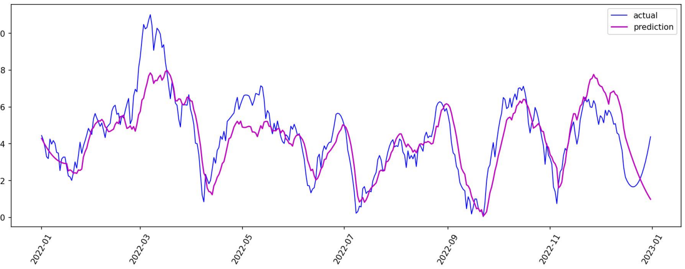

Figure 4. Random Forest's smoothed predicted (as computed by the best model) and actual values of daily hospitalisations for cardiovascular diseases for Brescia.

The value of MAE computed from the model whose input were just the two most important features (resulting from the best model) was 0.49. These features were the rolling mean and the day-of-the-month information.

4) Daily hospitalisations for respiratory diseases: We computed the ABE to use as the value to be improved: it was equal to 1.05.

The following reported values are the metrics computed for the best model coming from the optimisation (i.e., 991 trees):

- MAE = 1.05
- SMAPE = 33.7%
- R² Score = 0.22
- Mean of actual hospitalisations = 2.02
- Mean of predicted hospitalisations = 1.99
- Dispersion of actual hospitalisations = 1.48
- Dispersion of predicted hospitalisations = 0.79.

Figures 5 and 6 plot the actual (in blue) and predicted (in magenta) respiratory hospitalisations and their smoothed version, respectively.

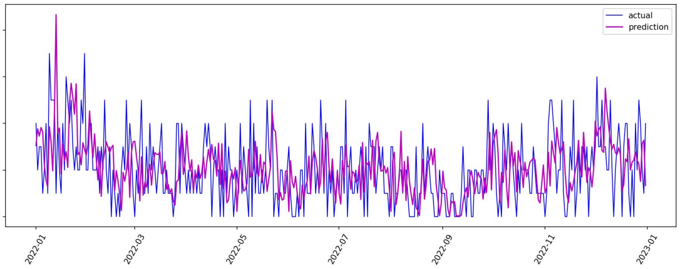

Figure 5. Random Forest's predicted (as computed by the best model) and actual values of daily hospitalisations for respiratory diseases for Brescia.

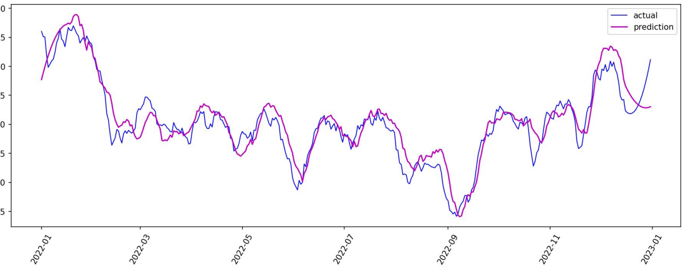

Figure 6. Random Forest's smoothed predicted (as computed by the best model) and actual values of daily hospitalisations for respiratory diseases for Brescia.

The value of MAE computed from the model whose input were just the two most important features (resulting from the best model) was 1.23. These features were the rolling mean and the day-of-the-month information.

BRESCIA'S PROVINCE

This section presents the analysis conducted on data of patients from the entire province of Brescia, which was never considered in the previous study [1].

5) Daily accesses' baseline: The computed value of ABE, the metric to be improved, was 12.79.

Again, here are reported the reached results for the 1000 trees model analysis of the baseline dataset:

- MAE = 10.49
- R² score = 0.51
- Acc* = 92.58%
- MAPE = 7.42%
- SMAPE = 3.69%
- Mean of actual accesses = 143.88
- Dispersion of actual accesses = 18.23.

6) Daily accesses: The following values are the metrics computed for the model created using the original number of trees (i.e., 1000):

- MAE = 9.81

- MAPE = 6.95%
- SMAPE = 3.45%
- Acc* = 93.05%
- R² Score = 0.55
- Mean of predicted accesses = 142.72
- Dispersion of predicted accesses = 12.56.

The following reported values are the metrics computed for the best model coming from the optimisation (i.e., 396 trees):

- MAE = 9.58
- MAPE = 6.81%
- SMAPE = 3.36%
- Acc* = 93.19%
- R² Score = 0.57
- Mean of predicted accesses = 143.15
- Dispersion of predicted accesses = 12.30.

Figures 7 and 8 plot the actual (in blue) and predicted (in magenta) accesses and their smoothed version, respectively.

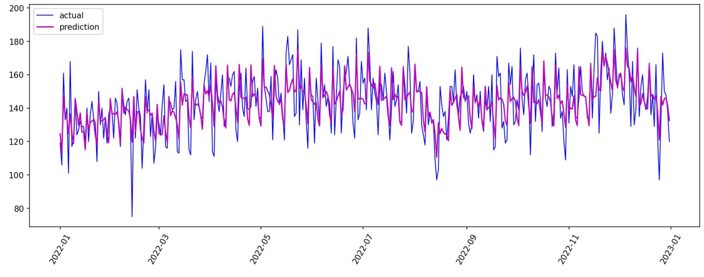

Figure 7. Random Forest's predicted (as computed by the best model) and actual values of daily ER accesses for the whole province of Brescia.

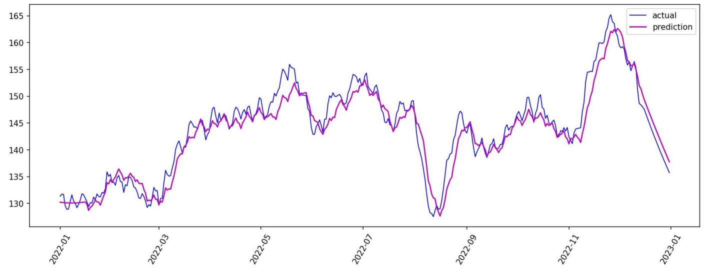

Figure 8. Random Forest's smoothed predicted (as computed by the best model) and actual values of daily ER accesses for the whole province of Brescia.

Computing a new best model (i.e., 940 trees), we tried a new approach. We divided the first four years of the dataset into the train and test portions casually rather than chronologically, and the obtained metrics were:

- MAE = 9.63
- R² Score = 0.74.

If we then used this same model to predict, as usual, the accesses for 2022 (as they represented completely new data for the algorithm), the metrics were:

- MAE = 9.84
- R² Score = 0.54.

The following metrics are those computed from the model whose input were just the two most important features (resulting from the best model). These features were the rolling mean and the Monday label.

- MAE = 12.78
- Acc* = 90.77%.

We were also interested to see which environmental variables were the most influential on the daily accesses, so we computed the best model and the feature importance for a dataset extracted from the original one without the rolling mean and on another where we removed the information about the days, too.

In the first case, the most important environmental variable was NO_x, while in the second case, the most important variables were NO_x, PM_{10}, RH_{max}, $\mathrm{PM}_{2.5}$ and T_{min}.

7) Daily hospitalisations for cardiovascular diseases: The computed value of ABE was 0.81.

The following reported values are the metrics computed for the best model coming from the optimisation (i.e., 1112 trees):

- MAE = 0.87
- SMAPE = 39.8%
- R² Score = 0.06
- Mean of actual hospitalisations = 1.37
- Mean of predicted hospitalisations = 1.38
- Dispersion of actual hospitalisations = 1.13
- Dispersion of predicted hospitalisations = 0.49.

Figures 9 and 10 plot the actual (in blue) and predicted (in magenta) cardiovascular hospitalisations and their smoothed version, respectively.

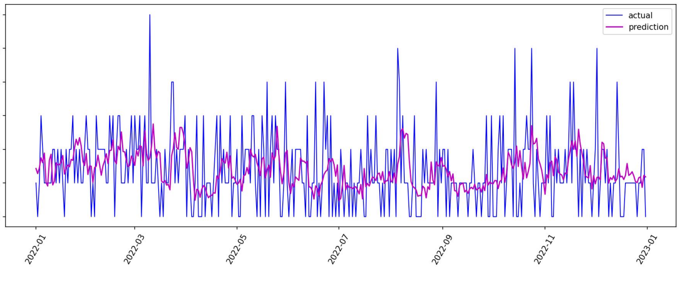

Figure 9. Random Forest's predicted (as computed by the best model) and actual values of daily hospitalisations for cardiovascular diseases for the whole province of Brescia.

We were also interested to see which environmental variables were the most influential on the daily cardiovascular hospitalisations, so we computed the best model and the feature importance for a dataset extracted from the original one without the rolling mean and on another where we removed the information about the days, too.

In the first case, the most important environmental variables were NO_x, RH_{max}, T_{min} and PM_{10}; in the second case they were the same, with the addition of $\mathrm{PM}_{2.5}$.

8) Daily hospitalisations for respiratory diseases: The computed value of ABE was 1.95.

The following reported values are the metrics computed for the best model coming from the optimisation (i.e., 105 trees):

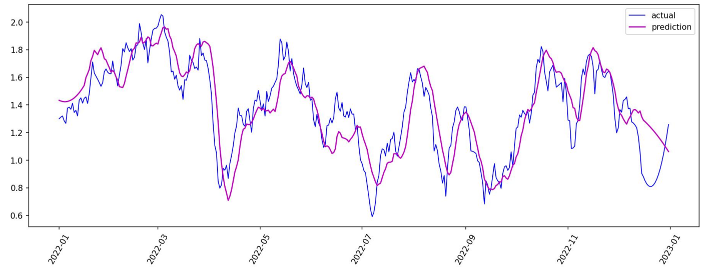

Figure 10. Random Forest's smoothed predicted (as computed by the best model) and actual values of daily hospitalisations for cardiovascular diseases for the whole province of Brescia.

- MAE = 1.96
- SMAPE = 14.3%
- R² Score = 0.45
- Mean of actual hospitalisations = 7.60
- Mean of predicted hospitalisations= 7.53
- Dispersion of actual hospitalisations = 3.35
- Dispersion of predicted hospitalisations = 2.37.

Figures 11 and 12 plot the actual (in blue) and predicted (in magenta) respiratory hospitalisations and their smoothed version, respectively.

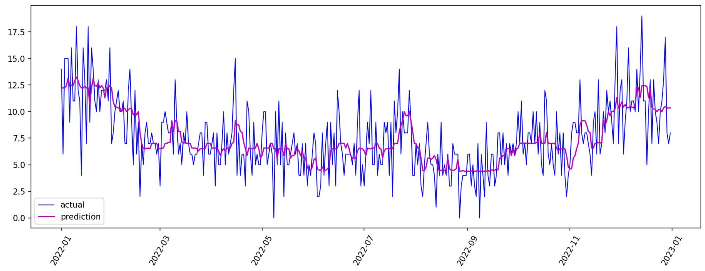

Figure 11. Random Forest's predicted (as computed by the best model) and actual values of daily hospitalisations for respiratory diseases for the whole province of Brescia.

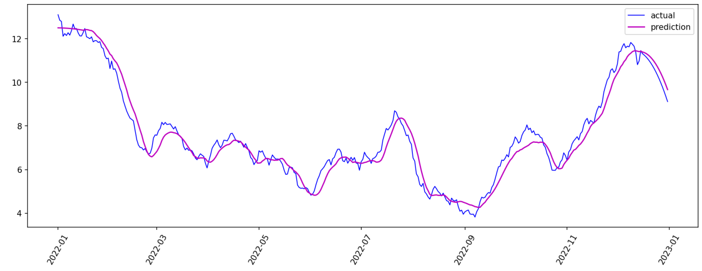

Figure 12. Random Forest's smoothed predicted (as computed by the best model) and actual values of daily hospitalisations for respiratory diseases for the whole province of Brescia.

We were also interested to see which environmental variables were the most influential on the daily respiratory hospitalisations, so we computed the best model and the feature importance for a dataset extracted from the original one without the rolling mean and on another where we removed the information about the days, too.

In the first case, the most important environmental variables were NO$_x$ and T$_{min}$, while in the second case, they were T$_{min}$, PM$_{2.5}$, NO$_x$, PM$_{10}$, O$_3$ and RH$_{max}$.

B. *Artificial Neural Network*

Here will be reported the metrics and plots resulting from the application (described in Subsection II-E) of a shallow 2-layer ANN to the hospitalisations caused by cardiovascular or respiratory disorders for patients coming both from only the city of Brescia and those from its entire province too.

Again, both numerical results of metrics and graphs are reported.

City of Brescia

1) Daily hospitalisations for cardiovascular diseases: The ABE to be improved was equal to 0.49, and the computed MAE for the ANN applied to the hospitalisations for cardiovascular diseases for Brescia was equal to 0.53. The SMAPE was 78.51%.

Figures 13 and 14 plot the actual (in green) and predicted (in magenta) Brescia's cardiovascular hospitalisations and their smoothed version, respectively.

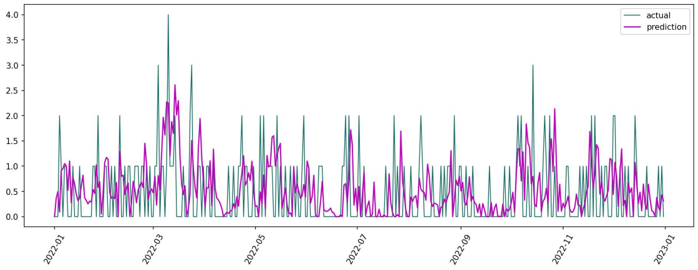

Figure 13. Artificial Neural Network's predicted and actual values of daily hospitalisations for cardiovascular diseases for Brescia.

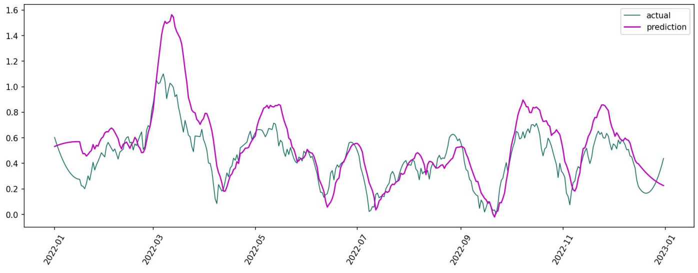

Figure 14. Artificial Neural Network's smoothed predicted and actual values of daily hospitalisations for cardiovascular diseases for Brescia.

2) Daily hospitalisations for respiratory diseases: The ABE to be improved was equal to 1.05, and the computed MAE for the ANN applied to the hospitalisations for respiratory diseases for Brescia was equal to 1.19. The SMAPE was 39.46%.

Figures 15 and 16 plot the actual (in green) and predicted (in magenta) Brescia's respiratory hospitalisations and their smoothed version, respectively.

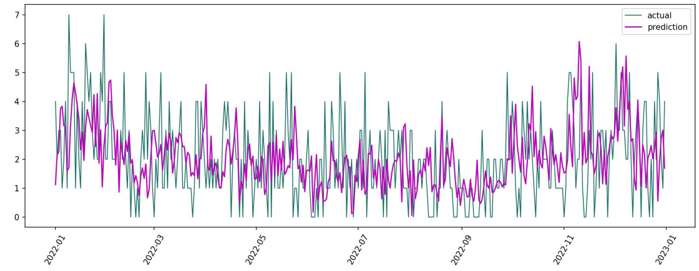

Figure 15. Artificial Neural Network's predicted and actual values of daily hospitalisations for respiratory diseases for Brescia.

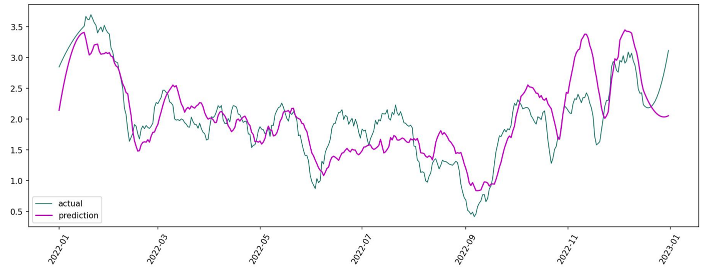

Figure 16. Artificial Neural Network's smoothed predicted and actual values of daily hospitalisations for respiratory diseases for Brescia.

Brescia's province

3) Daily hospitalisations for cardiovascular diseases: The ABE to be improved was equal to 0.81, and the computed MAE for the ANN applied to the hospitalisations for cardiovascular diseases for the province of Brescia was equal to 1.17. The SMAPE was 47.79%.

Figures 17 and 18 plot the actual (in green) and predicted (in magenta) cardiovascular hospitalisations and their smoothed version, respectively.

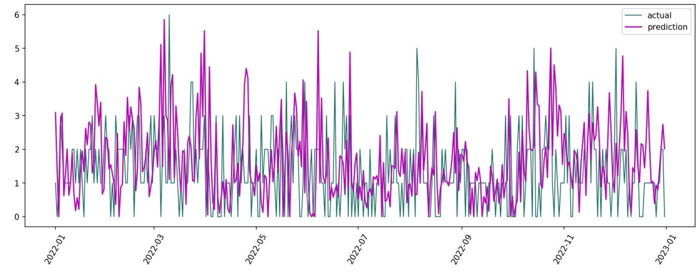

Figure 17. Artificial Neural Network's predicted and actual values of daily hospitalisations for cardiovascular diseases for the whole province of Brescia.

4) Daily hospitalisations for respiratory diseases: The ABE to be improved was equal to 1.95, and the computed MAE for the ANN applied to the hospitalisations for respiratory diseases for the province of Brescia was equal to 3.34. The SMAPE was 20.00%.

Figures 19 and 20 plot the actual (in green) and predicted (in magenta) respiratory hospitalisations and their smoothed version, respectively.

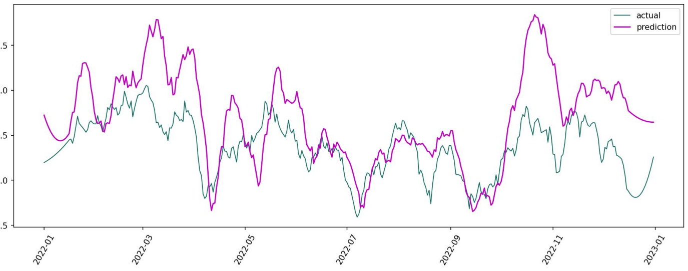

Figure 18. Artificial Neural Network's smoothed predicted and actual values of daily hospitalisations for cardiovascular diseases for the whole province of Brescia.

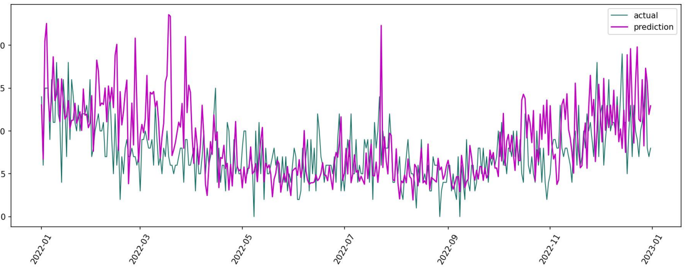

Figure 19. Artificial Neural Network's predicted and actual values of daily hospitalisations for respiratory diseases for the whole province of Brescia.

C. Support Vector Regression

Here are the results of the approach described in Subsection II-F to analyse and improve the predictions of daily hospitalisations for both cardiovascular and respiratory diseases for the city and province of Brescia.

As always, both numerical results of metrics and graphs are reported.

City of Brescia

1) Daily hospitalisations for cardiovascular diseases: The ABE to be improved was equal to 0.49, and the computed MAE for the SVR applied to the hospitalisations for cardiovascular diseases from Brescia was 0.50. The R² Score was 0.09.

Figures 21 and 22 plot the actual (in grey) and predicted (in magenta) cardiovascular hospitalisations and their smoothed version, respectively.

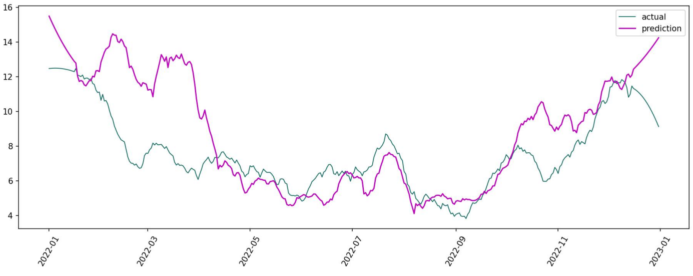

Figure 20. Artificial Neural Network's smoothed predicted and actual values of daily hospitalisations for respiratory diseases for the whole province of Brescia.

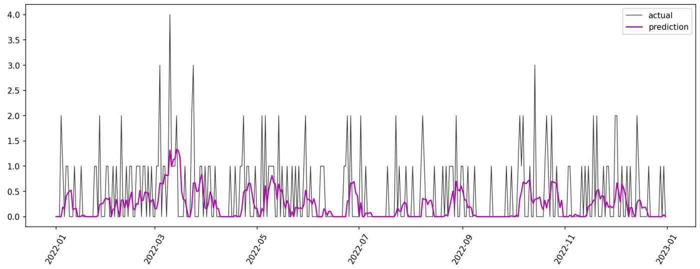

Figure 21. Support Vector Machine's predicted and actual values of daily hospitalisations for cardiovascular diseases for Brescia.

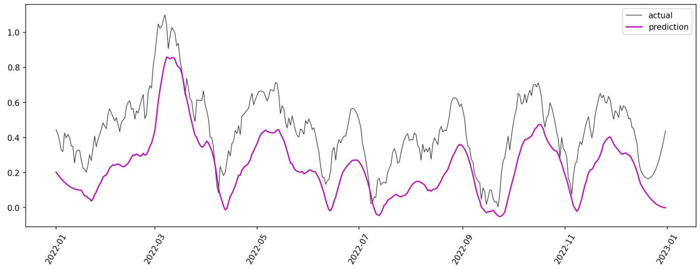

Figure 22. Support Vector Machine's smoothed predicted and actual values of daily hospitalisations for cardiovascular diseases for Brescia.

2) Daily hospitalisations for respiratory diseases: The ABE to be improved was equal to 1.05, and the computed MAE for the SVR applied to the hospitalisations for respiratory diseases from Brescia was 1.04. The R² Score was 0.39.

Figures 23 and 24 plot the actual (in grey) and predicted (in magenta) Brescia's respiratory hospitalisations and their smoothed version, respectively.

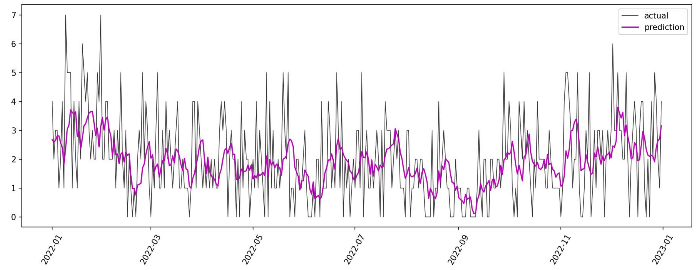

Figure 23. Support Vector Machine's predicted and actual values of daily hospitalisations for respiratory diseases for Brescia.

BRESCIA'S PROVINCE

3) Daily hospitalisations for cardiovascular diseases: The ABE to be improved was equal to 0.81, and the computed MAE for the SVR applied to the hospitalisations for cardiovascular diseases from Brescia was 0.83. The R² Score was 0.12.

Figures 25 and 26 plot the actual (in grey) and predicted (in magenta) cardiovascular hospitalisations and their smoothed version, respectively.

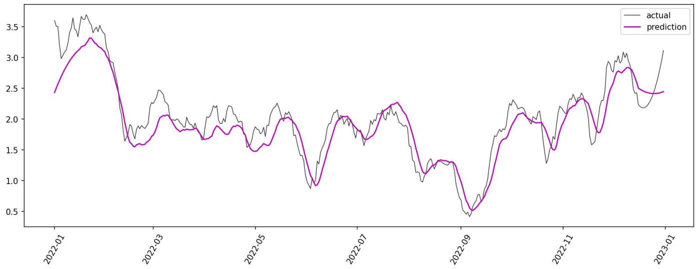

Figure 24. Support Vector Machine's smoothed predicted and actual values of daily hospitalisations for respiratory diseases for Brescia.

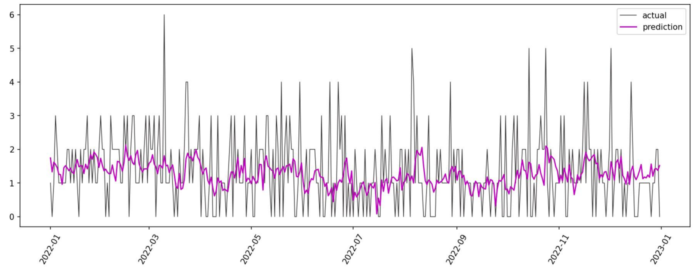

Figure 25. Support Vector Machine's predicted and actual values of daily hospitalisations for cardiovascular diseases for the whole province of Brescia.

4) Daily hospitalisations for respiratory diseases: The ABE to be improved was equal to 1.95, and the computed MAE for the SVR applied to the hospitalisations for respiratory diseases from Brescia was 1.94. The R² Score was 0.66.

Figures 27 and 28 plot the actual (in grey) and predicted (in magenta) respiratory hospitalisations and their smoothed version, respectively.

D. ARIMA

As already noted in Subsection II-G, here will be reported only one striking aspect of the daily hospitalisations for respiratory diseases of patients from Brescia.

Figure 29 plots the actual values coming from the train (2018-2021) portion of the respiratory hospitalisation dataset for the city, while Figure 30 the predictions (in magenta) of the 2022's test values (in brown).

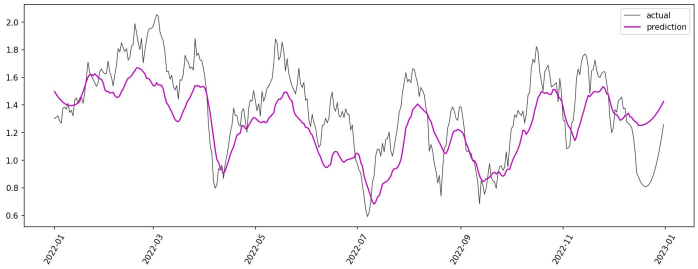

Figure 26. Support Vector Machine's smoothed predicted and actual values of daily hospitalisations for cardiovascular diseases for the whole province of Brescia.

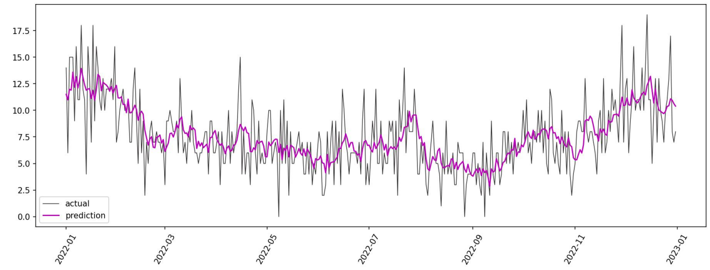

Figure 27. Support Vector Machine's predicted and actual values of daily hospitalisations for respiratory diseases for the whole province of Brescia.

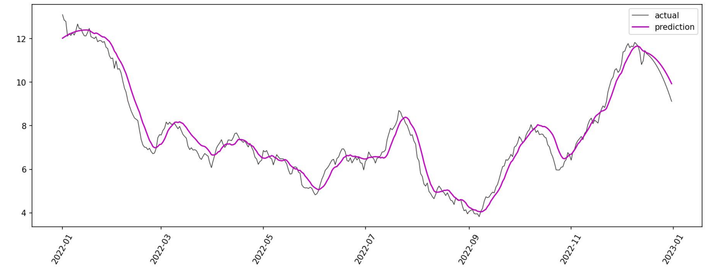

Figure 28. Support Vector Machine's smoothed predicted and actual values of daily hospitalisations for respiratory diseases for the whole province of Brescia.

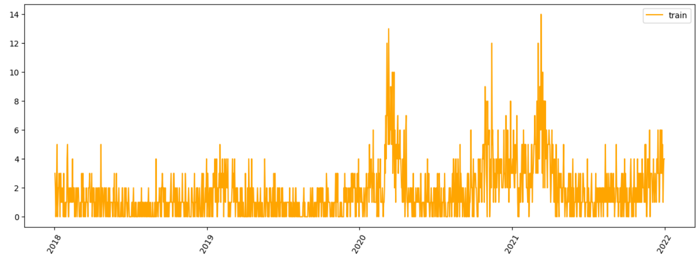

Figure 29. Actual daily hospitalisations caused by respiratory diseases for patients from Brescia from 2018 to 2021.

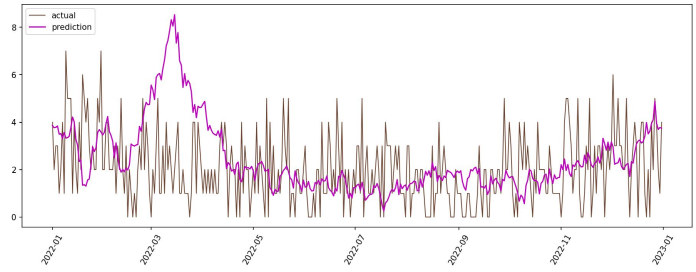

Figure 30. ARIMA's predicted and actual values of daily hospitalisations for respiratory diseases for Brescia.

IV. Discussion

The results, obtained applying the different predictive algorithms, reported in Section III will now be discussed.

A. *Random Forest*

The following are evaluations and comments on the results reported in Subsection III-A.

City of Brescia

1) Daily accesses: The results reported in Subsubsection III-A2, referring to the daily accesses of patients coming only from the city of Brescia, will now be discussed.

The error to improve (i.e., ABE) was 4.92. The achieved results for the baseline model represent the goodness of a poor prediction and can be used to evaluate if and how adding environmental data can improve the forecast.

The RF with the same number of estimators as the previous paper [1] already had better performances as its MAE was lower (i.e., 4.75).

The Acc* of this prediction, as computed from MAPE, was an appreciable 85.63%, and SMAPE was 6.9%. Since SMAPE accounts for the relative and balanced difference between predicted and actual values, such a low score indicates that the model predicts rather well.

Trying different approaches to evaluate how close the predictions were to the actual values, we computed their mean and dispersion. The obtained results confirm that, even though the dispersion is not as large, the mean is quite close, indicating that the general trend has been rightly forecasted.

The R² score (i.e., 0.21) was not high, but it was positive and, considering the complexity of the analysed scenario, can be deemed acceptable.

In an attempt to improve the performances further, the best model was computed over several trials, and the achieved MAE was even lower and equal to 4.63.

Even though the mean and dispersion of the actual and predicted values were not as good, our main objective was to minimise the MAE, and this model succeeded. Moreover, the Acc* was higher (i.e., 86.09%), so the MAPE was minor, and the SMAPE value was slightly too.

Unfortunately, the R² score was still not especially solid, as it was only 0.24.

The plots visually confirm the predictor's considerably satisfactory ability to follow the general trend.

Comparing these results with the analysis done on the dataset without environmental features, we can see how adding this type of information results in better metrics, thus producing a more precise forecast.

The metrics resulting from the best model applied to only the two most important features (none of which was environmental) were not as satisfying as the ones from the whole dataset, further proving that adding climate and pollution data influences the prediction positively.

2) Daily cardiovascular hospitalisations: The results reported in Subsubsection III-A3, referring to the daily hospitalisations due to cardiovascular diseases of patients coming only from Brescia, will now be discussed.

Since the model was, in this case, applied to sparse data, the Acc* could not be computed.

The error to improve was 0.49. The smaller MAE value was 0.51, obtained by applying the best model with 855 trees, and still worse than the ABE.

The R² score was too small (and the smallest yet) to be adequate, as it was too close to 0. Confirming this consideration, SMAPE, reaching 77.86%, was significantly greater than the others.

It appears clear that the approach needed to be changed to reach a better forecast of these accesses.

Regarding the mean and dispersion of the actual and predicted values, the former only had a 0.03 difference, while the latter had a 0.37 one. It means the forecast has a close but slightly narrower point cloud of accesses.

Still, the plots, especially the smoothed one, show that the model forecasts the general trend with acceptable precision.

The MAE resulting from the application only to the two most important features (none of which was environmental) was smaller than the one coming from the whole dataset but still not lower than ABE, so it cannot be considered successful.

3) Daily respiratory hospitalisations: The results reported in Subsubsection III-A4, referring to the daily hospitalisations due to respiratory diseases of patients coming only from Brescia, will now be discussed.

Since the model was, again, applied to sparse data, the Acc* could not be computed.

The error to improve was 1.05, and the best model (with 991 trees) achieved a MAE value equal to it. Since the value has remained equal and not lowered, the model cannot yet be considered satisfying.

R² score and SMAPE were not satisfying either: the approach needed to be changed to better forecast these accesses.

Regarding the mean and dispersion of the actual and predicted values, the former only had a 0.03 difference, and the latter a 0.69 one.

Even though the metrics are not ideal, the plots, especially the smoothed one, show that the general trend has been quite rightly forecasted.

The MAE resulting from the application on only the two most important features (none of which was environmental) was even higher than the one coming from the whole dataset, which was merely equal to the ABE. It further proves that adding climate and pollution data influences the predictor performances positively.

BRESCIA'S PROVINCE

4) Daily accesses: The results reported in Subsubsection III-A6, referring to the daily accesses of patients from the entire province of Brescia, will now be discussed.

The error to improve was 12.79. The RF with the same number of estimators as the previous paper already had better performances as its MAE was lower (i.e., 9.81).

The Acc* of this prediction, as computed from MAPE, was a satisfying 93.05%.

Trying different approaches to evaluate how close the predictions were to the actual values, we computed their mean and dispersion. The obtained results confirm that, even though the dispersion is not as large, the mean is quite close, indicating that the general trend has been rightly forecasted.

Even if the R² score was only 0.55, the model seems valid, and SMAPE was only 3.45%, so the prediction's errors are reasonably negligible, resulting in the best values reached for these metrics yet, thus the most precise model.

In trying to improve these performances further, the best model was computed over several trials, and the achieved metrics were, in fact, even better.

The MAE, equal to 9.58, was even lower, and the Acc* of the prediction (i.e., 93.19%) was slightly higher.

Regarding the mean and dispersion of the actual and predicted values, the first one was even closer, while the dispersion marginally worsened. Still, the general trend has been rightly forecasted.

Even if the R² score was only 0.57, it is still the best achieved one, considering all previous case studies, while SMAPE was the smallest one, as it was equal to only 3.36%.

The plots visually confirm the predictors' ability to follow the general trend.

When trying to divide the train and test datasets casually instead of chronologically, the metrics appeared to be better as we reached, through its own best model, an R² Score as high as 0.74.

For this reason, we tried further using this different approach, but we also had to test it on future chronologically presented data, as that is how future input data would look.

Unfortunately, though, when tested on 2022 data, the model performance returned to values closer to the ones from the initial chronological division. So, we decided to discard this plan and revert to the original one.

The metrics resulting from applying the best model on only the two most important features (none of which was environmental) were not as satisfying as the ones from the whole dataset, proving further that adding climate and pollution data influences the prediction positively.

Regarding the influence of environmental variables on ER accesses, it is interesting to note that low temperatures and humidity rates hold this much of an impact, as some polluting substances do. The fact that minimum temperature and pollution substances appear together is unsurprising since heating systems release pollutants like $PM_{2.5}$ and PM_{10}.

5) Daily cardiovascular hospitalisations: The results reported in Subsubsection III-A7, referring to the daily hospitalisations due to cardiovascular diseases of patients from the entire province of Brescia, will now be discussed.

Since, for this case study, the model was applied again to sparse data, the Acc* could not be computed.

The error to improve was 0.81. The lower MAE value was 0.87, through the application of the best model having 1112 trees, nevertheless worse than the ABE.

The R² score (i.e., 0.06) was even smaller than the city (reported in Subsubsection III-A3) one, even if the SMAPE (i.e., 39.8%) was minor. This worse R² score could be due to the added sparsity of data from adding patients that follow the same noisy general trend (way different than the whole accesses' one).

Clearly, the approach needed to be changed to forecast these accesses better.

Regarding the mean and dispersion of the actual and predicted values, the former only had a 0.01 difference, while the latter had a 0.64 one. It means that the forecast has a narrower point cloud of accesses.

Still, the plots, especially the smoothed one, show that the model forecasts the general trend with acceptable precision.

Regarding the influence of environmental variables on cardiovascular hospitalisations coming from triage, it is interesting to note that humidity and temperature have such an impact, along with some polluting substances.

6) Daily respiratory hospitalisations: The results reported in Subsubsection III-A8, referring to the daily hospitalisations due to respiratory diseases of patients from the entire province of Brescia, will now be discussed.

Since the model was, again, applied to sparse data, the Acc* could not be computed.

The error to improve was 1.95. The best model (with 105 trees) achieved a MAE value of 1.96, still slightly higher than the ABE, an R² score of 0.45, and a SMAPE of 14.3%. Even though these last two values were better than the ones reached for the city's respiratory hospitalisations and less unsatisfactory than the MAE, the approach still needed to be changed to achieve a valid forecast.

Regarding the mean and dispersion of the actual and predicted values, the former only had a 0.07 difference, and the latter a 0.98 one.

Still, the plots, especially the smoothed one, show that the general trend has been satisfyingly forecasted.

Regarding the influence of environmental variables on respiratory hospitalisations coming from triage, it is interesting to note that minimum temperature has such an impact, along with more polluting substances (compared with the other case studies). It was expected because of evidence that respiratory disorders flares link to air pollution [9] [19].

B. Artificial Neural Network

The following are evaluations and comments on the results reported in Subsection III-B.

City of Brescia

1) Daily cardiovascular hospitalisations: The results reported in Subsubsection III-B1, referring to the daily hospitalisations of patients affected by cardiovascular diseases coming from Brescia, will now be discussed.

The ABE to improve was 0.49, but, unfortunately, the network's MAE (i.e., 0.53) was higher, even more than the RF one.

SMAPE was 78.51%, again, worse than the RF one.

It further proved that the best predictive algorithm approach for this analysis was yet to be found. The plots do not appear remarkably different from the RF ones, but, nevertheless, not as good as them.

2) Daily respiratory hospitalisations: The results reported in Subsubsection III-B2, referring to the daily hospitalisations of patients affected by respiratory diseases coming from Brescia, will now be discussed.

The ABE to improve was 1.05, but, unfortunately, the network's MAE (i.e., 1.19) was higher and even more than the RF one.

SMAPE was 39.46%, again, worse than the RF one.

It further proved that, even if the plots do not appear tragically different from the actual values, the best predictive algorithm approach for hospitalisations was yet to be found.

Brescia's province

3) Daily cardiovascular hospitalisations: The results reported in Subsubsection III-B3, referring to the daily hospitalisations of patients affected by cardiovascular diseases coming from the entire province of Brescia, will now be discussed.

The ABE to improve was 0.81, but the network's MAE (i.e., 1.17) was still unsatisfactory and worse than the RF one. Same for SMAPE as it was higher.

It deeply proved that the best predictive algorithm approach for this analysis was yet to be found, as the plots appear clearly different and worse than the RF ones.

4) Daily respiratory hospitalisations: The results reported in Subsubsection III-B4, referring to the daily hospitalisations of patients affected by respiratory diseases from the entire province of Brescia, will now be discussed.

The ABE to improve was 1.94, and the network's MAE (i.e., 3.34) was absolutely unsatisfactory and way worse than the RF one. SMAPE was surprisingly small as it was equal to 20%, but still higher than the RF one.

It further proved that the best predictive algorithm approach for this hospitalisation analysis was yet to be found, as the plots appear to diverge from the actual values significantly.

C. Support Vector Machine

The following are evaluations and comments on the results reported in Subsection III-C.

Since ANN did not improve as hoped, we approached another algorithm, finally obtaining better results for one of the two hospitalisations' groupings.

City of Brescia

1) Daily cardiovascular hospitalisations: The results reported in Subsubsection III-C1, referring to the daily hospitalisations due to cardiovascular diseases of patients coming only from Brescia, will now be discussed.

The ABE was 0.49, and the reached MAE was 0.50. Even if it is not better than the baseline error, it is still slightly an improvement, compared to the RF error.

Instead, the R² score was dramatically lower because it was 0.09.

By visually analysing the plots, it can be commented that the SVR prediction underestimates the daily hospitalisations.

2) Daily respiratory hospitalisations: The results reported in Subsubsection III-C2, referring to the daily hospitalisations due to respiratory diseases of patients coming only from Brescia, will now be discussed.

The ABE was 1.05, and the reached MAE was 1.04. It represents a case study where the application of a different model did, indeed, improve performances.

Further proving this point, the RF's R² score was only 0.22, while SVR's was 0.39.

The plots appear way more adherent, too, resulting in a satisfying forecast of a notably complex application.

Brescia's province

3) Daily cardiovascular hospitalisations: The results reported in Subsubsection III-C3, referring to the daily hospitalisations due to cardiovascular diseases of patients coming only from the entire province of Brescia, will now be discussed.

The ABE was 0.81, and the reached MAE was 0.83. Even if it is not better than the baseline error, it is still an improvement compared to the RF one.

The same goes for the R² score since it even doubled.

Compared with the RF plots, these appear less adherent to actual data, and they are slightly underestimating.

4) Daily respiratory hospitalisations: The results reported in Subsubsection III-C4, referring to the daily hospitalisations due to respiratory diseases of patients coming only from the whole province of Brescia, will now be discussed.

The ABE was 1.95, and the reached MAE was 1.94. Again, this represents another time when applying a predictive model improved performances. Even the best RF model did not obtain a MAE value smaller than ABE.

Further proving this point, the R² score was a striking 0.66, the highest value of this metric we reached in any trial, as we discarded the non-chronological approach.

The plots appear way more adherent, too, especially the smoothed one.

It resulted in the best forecast of all, even though we must highlight that we have not applied SVR to daily accesses as we had already found valid models, so we do not know which results would have come out of that.

D. ARIMA

The following are evaluations and comments on the results reported in Subsection III-D.

Based on the previous findings [1], we already knew that ARIMA was not the ideal model to improve the performances of our forecast, but we still decided to run it to see if we could find any aspect of interest.

Since ARIMA is a time-series-based analysis, trend fluxes heavily influence it: this resulted in a peculiar prediction graph for respiratory diseases-caused hospitalisations of patients coming from Brescia, as it predicted a phantom positive peak around March.

As we investigated the reason for that, we found its explanation in the observable trend of the previous years' actual data: in fact, March 2020 and 2021 saw a surge in hospitalisations due to respiratory disorders as more patients contracted COVID-19.

The main drawback of time series models is that they rely only upon the forecasted variable without comprehending and looking for the concealed causes of its behaviour. Still, they can represent a suitable approach when dealing with real-life daily chronological data.

V. Conclusion and Future Work

When analysing metrics and graphs from the different models, we can appreciate how, in the end, for both the city of Brescia and its province, we could manage to validly predict daily accesses and hospitalisations due to respiratory diseases.

The same cannot be said for cardiovascular hospitalisations, plausibly due to the high sparsity of these data, meaning that further research needs to be undertaken. Note that the number of hospitalisations for specific pathologies is limited to a few people every day and, sometimes, even none, and this is particularly noticeable for cardiovascular disorders.

Still, the main objective of this work, which was to upgrade and deepen the previously reported analysis [1], was generally reached. Even the worst result, coming from the analysis of cardiovascular hospitalisations of patients, still represents an improvement from the previous study, and the latter's findings have been validated.

Focusing on comparing the different predictive algorithms, we can state that, for these specific datasets, SVR seems to be the best one, followed by RF. ANN, instead, results in performances closer to the ones of ARIMA.

Visually analysing the plots, our best forecasts of daily accesses and respiratory hospitalisations appear to adhere quite well to the actual data, and their metrics are quite satisfying, too.

In fact, generally speaking, even if the specific values are not always correctly predicted, the overall trend seems to be rightly followed, and peak values (like surges in accesses or hospitalisations) are captured.

Another significant result to highlight is the confirmation of how adding environmental data can improve the prediction.

When we tried to apply the same models to reduced versions of the datasets that only contained calendrical information or, instead, discarded it, we generally achieved better performances.

Based on these observations, this work represents a coherent deep-dive that further analyses the previous approach.

The prediction of ER accesses and hospitalisations from a specific geographical area through the analysis of clinical and environmental data is feasible.

The previous promising results have been confirmed and improved, even if this method's application on cardiovascular hospitalisations could still benefit from further investigation.

Nevertheless, we cannot generalise the results since we obtained them by analysing a period majorly made up of

COVID-19-ridden years and a limited geographical area. Thus, we can only use them to comment on this specific frame.

The performances could dramatically differ if the analogous pre-processing and the same models were applied to other contexts or just even on a longer and more stable period.

In summary, our hypothesis of enabling forecasting of ER volumes by combining historical clinical, weather and pollution data, linked by a detailed geographical indication, has been proven to be suitable and also given more than encouraging results.

Although additional work could still be encouraged to improve the achieved performances, this represents a new point of view on such a complex and poignant matter.

The real-life application of this approach is now possible, and its adaptation to other areas appears simple, even if we cannot predict how accurate that forecast would be.

To conclude, future developments of this work will widen to other areas, with the hope of moving to ever-growing datasets, and additional algorithm testing will be conducted to improve the best-achieved predictions further.

Nevertheless, any additional attempt will gather supplementary valuable insight on this topic and shed light on how our surrounding environment influences human health.

This One Health approach may offset a new way of managing ER worldwide, enabling the monitoring of entire populations and geographical areas, with the final objective of improving the quality of healthcare and people's quality of life.

References

[1] I. Della Torre, I. Avellino, F. Marinaro, A. Buccoliero, and A. Colangelo, "Predictive analytics for Emergency Department visits based on local short-term pollution and weather exposure", AIHealth 2024, The First International Conference on AI-Health. ThinkMind, pp. 29-34, 2024.

[2] J. D. Sonis and B. A. White, "Optimizing patient experience in the emergency department", Emergency Medicine Clinics, vol. 38, no. 3, pp. 705–713, 2020.

[3] Z. Qiao et al., "Using machine learning approaches for emergency room visit prediction based on electronic health record data", Building continents of knowledge in Oceans of data: The future of co-created eHealth. IOS Press, pp. 111–115, 2018.

[4] Y. M. Chiu, J. Courteau, I. Dufour, A. Vanasse, and C. Hudon, "Machine learning to improve frequent emergency department use prediction: a retrospective cohort study", Scientific Reports, vol. 13, no. 1, p. 1981, 2023.

[5] C. Peláez-Rodríguez, R. Torres-López, J. Pérez-Aracil, N. López-Laguna, S. Sánchez-Rodríguez, and S. Salcedo-Sanz, "An explainable machine learning approach for hospital emergency department visits forecasting using continuous training and multi-model regression" Computer Methods and Programs in Biomedicine, vol. 245, 2024.

[6] A. Cameron, K. Rodgers, A. Ireland, R. Jamdar, and G. A. McKay, "A simple tool to predict admission at the time of triage", Emergency Medicine Journal, vol. 32, no. 3, pp. 174–179, 2015.

[7] R. Sánchez-Salmerón et al., "Machine learning methods applied to triage in emergency services: A systematic review", International Emergency Nursing, vol. 60, 2022.

[8] W. Zhu et al., "The effect and prediction of diurnal temperature range in high altitude area on outpatient and emergency room admissions for cardiovascular diseases", International Archives of Occupational and Environmental Health, vol. 94, no. 8, pp. 1783–1795, 2021.

[9] T. Abe et al., "The relationship of short-term air pollution and weather to ED visits for asthma in Japan", The American journal of emergency medicine, vol. 27, no. 2, pp. 153–159, 2009.

[10] D. Martinaitiene and N. Raskauskiene, "Weather-related subjective well-being in patients with coronary artery disease", International Journal of Biometeorology, vol. 65, pp. 1299–1312, 2021.

[11] M. Hensel et al., "Association between weather-related factors and cardiac arrest of presumed cardiac etiology: a prospective observational study based on out-of-hospital care data", Prehospital Emergency Care, vol. 22, no. 3, pp. 345–352, 2018.

[12] S. Kojima et al., "Fine particulate matter and out-of-hospital cardiac arrest of respiratory origin", European Respiratory Journal, vol. 57, no. 6, p. 2004299, 2021.

[13] M. A. Shahrbaf, M. A. Akbarzadeh, M. Tabary, and I. Khaheshi, "Air pollution and cardiac arrhythmias: a comprehensive review", Current Problems in Cardiology, vol. 46, no. 3, p. 100649, 2021.

[14] J. M. Delgado-Saborit et al., "A critical review of the epidemiological evidence of effects of air pollution on dementia, cognitive function and cognitive decline in adult population", Science of the Total Environment, vol. 757, p. 143734, 2021.

[15] S.-T. Zang et al., "Ambient air pollution and COVID-19 risk: evidence from 35 observational studies", Environmental research, vol. 204, p. 112065, 2022.

[16] M.-Y. Wu, W.-C. Lo, C.-T. Chao, M.-S. Wu, and C.-K. Chiang, "Association between air pollutants and development of chronic kidney disease: a systematic review and meta-analysis", Science of the Total Environment, vol. 706, p. 135522, 2020.

[17] Y. Li, L. Xu, Z. Shan, W. Teng, and C. Han, "Association between air pollution and type 2 diabetes: an updated review of the literature", Therapeutic Advances in Endocrinology and Metabolism, vol. 10, pp. 1-15, 2019.

[18] R. D. Brook et al., "Particulate matter air pollution and cardiovascular disease: an update to the scientific statement from the American Heart Association", Circulation, vol. 121, no. 21, pp. 2331–2378, 2010.

[19] F. Dominici et al., "Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases", Jama, vol. 295, no. 10, pp. 1127–1134, 2006.

[20] R. Duan et al., "Association between short-term exposure to fine particulate pollution and outpatient visits for ulcerative colitis in Beijing, China: A time–series study", Ecotoxicology and Environmental Safety, vol. 214, p. 112-116, 2021.

[21] F. Jaime et al., "Solar radiation is inversely associated with inflammatory bowel disease admissions", Scandinavian journal of gastroenterology, vol. 52, no. 6-7, pp. 730–737, 2017.

[22] C.-L. Chan et al., "A survey of ambulatory-treated asthma and correlation with weather and air pollution conditions within Taiwan during 2001–2010", Journal of asthma, vol. 56, no. 8, pp. 799–807, 2019.

[23] J. Lu et al., "Feasibility of machine learning methods for predicting hospital emergency room visits for respiratory diseases", Environmental Science and Pollution Research, vol. 28, pp. 29701–29709, 2021.

[24] https://civile.asst-spedalicivili.it/servizi/unitaoperative/unitaoperative_fase02.aspx?ID=586 [Retrieved online: November 2024].

[25] F. Tartari, A. Guglielmo, F. Fuligni, and A. Pileri, "Changes in emergency service access after spread of COVID-19 across Italy", Journal of the European Academy of Dermatology and Venereology, vol. 34, no. 8, p. e350, 2020.

[26] T. Ferrari, C. Zengarini, F. Bardazzi, and A. Pileri, "In-depth, single-centre, analysis of changes in emergency service access after the spread of COVID-19 across Italy", Clinical and Experimental Dermatology, vol. 46, no. 8, pp. 1588–1589, 2021.

[27] https://www.hypermeteo.com/ [Retrieved online: November 2024].

[28] https://www.istat.it/en/classification/codes-of-italian-municipalities-provinces-and-regions/ [Retrieved online: November 2024].

[29] https://www.arpalombardia.it/ [Retrieved online: November 2024].

[30] https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health [Retrieved online: November 2024].

[31] https://www.kaggle.com/datasets/nicholasjhana/energy-consumption-generation-prices-and-weather [Retrieved online: November 2024].

[32] S. Khomenko et al. "Premature mortality due to air pollution in European cities: a health impact assessment", The Lancet Planetary Health, vol. 5, no. 3, pp. e121–e134, 2021.

[33] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html [Retrieved online: November 2024].

[34] https://towardsdatascience.com/random-forest-in-python-24d0893d51c0 [Retrieved online: November 2024].

[35] https://optuna.org/ Retrieved online: November 2024].

[36] G. Vishwakarma, A. Sonpal, and J. Hachmann, “Metrics for benchmarking and uncertainty quantification: Quality, applicability, and best practices for machine learning in chemistry”, Trends in Chemistry, vol. 3, no. 2, pp. 155-156, 2021.

[37] J. Zou, Y. Han, and S. S. So, “Overview of artificial neural networks”, Artificial neural networks: methods and applications, pp. 14-22, 2009.

[38] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html [Retrieved online: November 2024].

[39] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, “Support vector machines”, IEEE Intelligent Systems and their Applications, vol. 13, no. 4, pp. 18-28, July-Aug. 1998.

[40] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html [Retrieved online: November 2024].

[41] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html [Retrieved online: November 2024].

[42] R. H., Shumway, and D. S., Stoffer, “ARIMA models”, Time series analysis and its applications: with R examples, pp 75-163, 2017.

[43] https://alkaline-ml.com/pmdarima/modules/generated/pmdarima.arima.auto_arima.html [Retrieved online: November 2024].

[44] https://www.statsmodels.org/stable/generated/statsmodels.tsa.arima.model.ARIMA.html [Retrieved online: November 2024].