Air Quality Prediction Using Machine Learning Methods: A Case Study of Bjelave Neighborhood, Sarajevo, BiH

Emina Džaferović and Kanita Karadžović-Hadžiabdić

Abstract Air pollution is a complex mixture of toxic components that has the direct impact on human health, life quality, and the environment. In this study, meteorological variables and concentration of air pollutants are used to predict the common air quality index (CAQI) in Bjelave neighborhood, Sarajevo, BiH. CAQI prediction models were built using five popular machine learning techniques in the air pollution domain: Support Vector Regression, Random Forest, Extreme Gradient Boosting, Multiple Linear Regression and Multilayer Perceptron, using three-year period data (2016–2018). Prediction performance was measured using regression metrics: R-squared and RMSE. Ensemble technique, Random Forest method achieved the best performance results from the five evaluated machine learning methods: $R^2 = 0.99$ and $RMSE = 2.30$, using the dataset when missing values were removed, and $R^2 = 0.99$ and $RMSE = 2.58$ using the dataset when missing values were imputed using linear regression method.

Keywords Industry 4.0 · Upgrading legacy machines · Machine vision

1 Introduction

Air pollution emerges as gases and particulate pollutants are introduced into the atmosphere by either natural processes or human activities. It is a large-scale problem around the globe in both developing and developed countries [1]. Fast pace demographic growth and giant energy requirements have resulted in the emission of air pollutants at a toxic level that affects human health and the surrounding environment.

According to the World Health Organization (WHO), air pollution in cities, as well as rural areas, causes approximately 4.2 million deaths per year due to stroke, heart disease, lung cancer, and chronic respiratory diseases, globally [2]. Some developed
regions such as California [3], have succeeded in decreasing air pollution by setting mechanisms like catalytic converters on cars and introducing stricter regulations on emission of particulate pollutants from sources like diesel engines in place, while others such as Mexico City and Beijing still remain somber. Even though this problem does affect developing as well as developed countries in the same way, the highest-burden is placed on low- and middle-income countries. According to [3], the Iranian city of Zabol that has a population of fewer than 150,000 people, has the world’s worst concentration of PM2.5 pollution due to the dust that is being generated as the surrounding wetlands are desiccated.

The common air quality index (CAQI) was developed by the European Union in 2006 with primary goals to develop the awareness of urban air pollution, its main source (traffic) and to make air quality comparable between different cities in Europe [4]. Accordingly, CAQI has no direct link to short-term health effects but rather assists in decision-making processes related to pollution measures taken by authorities.

When discussing air pollution and its sources, it is important to mention that the increase in the concentration of air pollutants in the atmosphere does not occur as a result of a sudden increase in emission, but rather from certain meteorological conditions that either impede dispersion in the atmosphere or increase the generation of the pollutant [5]. Meteorological conditions play an important role in air pollution due to the direct or indirect effect that they have on emissions, transport, formation, and deposition of air pollutants. As specified in [6], several studies have shown that meteorological factors such as wind speed and direction, temperature and relative humidity can significantly affect air quality. Above mentioned factors are some of the key considerations that need to be addressed when building the forecasting model.

The problem of air pollution is an issue that has been around for at least 50 years. However, over the past decade, the emergence of new technologies such as sensors led to a shift in air pollution estimation and monitoring. More specifically, the low cost of these devices resulted in a wider utilization of information by communities all over the world [7]. Furthermore, the advances in research led to a discovery that air pollution is a problem associated not only to a negative environmental effect but also to many health issues [8]. As a result, in the last couple of years just like in many cities worldwide, the awareness of the air pollution increased in Sarajevo and other cities in Bosnia and Herzegovina. Particles such as PM$_{10}$ (40 $\mu$g/m$^3$) and PM$_{2.5}$ (25 $\mu$g/m$^3$) have exceeded their average margin values which caused an increase in health issues [9].

Considering all of the factors and the impact that air pollution has on the environment and human health, it is of great importance to not only deal with the current situation but also to prevent future air pollution episodes. Short-term forecasting of air quality is important so that authorities can take preventive or warning measures during the episodes of high concentration of air pollutants in order to protect public. Impacting daily habits of population or placing restrictions on traffic and/or industry, decrease in excessive medication, need for hospital treatment and premature death should be achievable.

This study is the first in Bosnia and Herzegovina that performs regression based air quality prediction using five of the most popular machine learning techniques used
Air Quality Prediction Using Machine Learning Methods …

in and also outside of the air pollution domain: Support vector machine—regression (SVR), Random Forest (RF), XGBOOST, Multiple linear regression (MLR) and Multilayer perceptron (MLP). Furthermore, air pollution trend in the region of Bjelave neighborhood is analyzed using CAQI, making an important step in standardization with the rest of the European cities. According to the recent study [10], there is a strong link between air pollution and high death rates in people with COVID-19, which is why we hope to further raise the awareness of air pollution.

2 Related Work

Higher air quality index caused by increase in air pollution has been a hot topic for research in the past couple of years all over the world. The reason for this is the effect that air pollution has on the environment and particularly on human health. Due to the nature of the domain which consists of large volume of data and very complex and unknown deterministic equations, machine learning techniques have been a popular approach for prediction of air pollution throughout the world [11–14]. Most of the researches performed were for predicting concentration of a single pollutant, usually particulate matter due to the fact that it can have a devastating effect on human health [15]. It is often not clear beforehand which machine learning technique is the most suitable and gives the best prediction performance for the domain which is why different researches provide evaluation on different machine learning techniques.

For instance, authors of [13] have evaluated performance of four machine learning methods of different complexity in predicting the concentration of PM$_{10}$ for the region of Helsinki in Finland. Methods that they have chosen for this task were logistic regression, decision tree, multivariate adaptive regression splines (MARS) and neural network. The end result proved that the performance of three of four methods (logistic regression, MARS and neural network) was similar and was sensitive to the size of the learning sample and the time period used.

On the other hand, some of the papers such as [14] implement an innovative threefold intelligent hybrid system, HISYCOL, which was developed by combining several machine learning methods. Accuracy was improved by using unsupervised machine learning to cluster the data vectors and trace any type of hidden knowledge after which a Mamdani fuzzy inference system was employed for each air pollutant. Results show that generally speaking by paying attention to average values, the hybrid approach which was developed performs more or less in the same way that holistic approaches of robust adaptive multi-feedback (RAF) and feedforward neural network (FFNN) do.

While there has been much research like those previously mentioned which were concerned with prediction of concentration of air pollutants, few researchers have taken air quality index into consideration [16,17]. Authors of [16] performed principal component analysis (PCA) to identify the sources of air pollution and tree based ensemble learning models to predict the urban air quality index for the region of Lucknow in India. Machine learning methods that were used during the study
included single decision tree (SDT), decision tree forest (DTF), decision treeboost (DTB) and support vector machines (SVM). End results show that both DTF and DTB models outperformed SVM in both regression and classification tasks and the reasoning behind could be the usage of bagging and boosting algorithms in these models. On the other hand, paper [17] deals with real-time air quality index prediction. Results show that researchers were able to get the overall classification accuracy above 90% for all of the methods, having the neural network as the best performing method with an accuracy of 99.56%.

The above-mentioned studies as well as other studies not covered in this literature review, from the field of machine learning and air pollution, show that a lot of effort has been made in terms of air quality prediction. Machine learning methods have been widely used in environmental science for solving many problems. In terms of air pollution, we could see that the most popular methods used include SVM, MLP, MLR, DTs and ANN. Accordingly, this paper provides comparison between some of the most popular machine learning algorithms used in and outside of the current domain.

3 Materials and Methods

3.1 Dataset

Sarajevo is a city in a central part of Bosnia and Herzegovina that is spread across the area of 141.5 km² with the average elevation of approximately 500 m above sea level. The city has a humid continental climate¹ and four seasons with uniformly spread precipitation throughout the year. The average yearly temperature is 10 °C, with January being the coldest month and July being the warmest month during the year.

The data was obtained from the Federal Hydrometeorological Institute BiH (FHMZ).

Even though Sarajevo has five measuring stations, this study uses the data collected from Bjelave station only, since at the time when the research was performed, needed data from other meteorological stations in Sarajevo was not accessible. Bjelave station is located at 18° 25′ 22.94″ and 43° 52′ 2.40″ longitude and latitude (Fig. 1), and measures hourly values for both concentration of air pollutants and meteorological variables. The obtained dataset consists of hourly values of five air pollutant concentrations (PM₁₀, NO₂, SO₂, O₃ and CO) whose statistical analysis is given in Table 1, and eight meteorological variables (minimum, maximum and average temperature, wind speed, wind direction, humidity, pressure and precipitation) gathered from 2016 to 2018 (three-year period).

¹Climatic region, defined by climatologist Wladimir Köppen in 1900, that is characterized by large seasonal temperature difference.
Before using the obtained data as inputs into the machine learning methods, handling of missing values and data normalization was done as a preprocessing step. Malfunctions in data collection are said to be of either partial deficiency, where either the measuring station malfunctions for a shorter period of time (up to several months), or total deficiency where the measuring station malfunctions for a longer period of time (up to several years) [14]. Obtained Bjelave dataset contained randomly spread partial deficiencies which means that some of the data was missing for a couple of days up to a couple of months. Missing air pollutant concentration data was imputed using a linear interpolation technique based on the work of Noor et al. [19]. In this work, the authors found that imputation of missing values using linear interpolation achieves the best results when missing value percentage is between 5 and 15% in air pollution related problems. Figure 2 shows the percentages of missing values for each feature used in the dataset. The figure shows that in all cases but one, (i.e. PM$_{10}$—17%) the percentage of missing values are within the required boundaries.

Additionally, concentrations of air pollutants together with meteorological variables were normalized in order to avoid the problem of having features with a wider range (e.g. CO) dominate over the features with a narrower range (e.g. PM$_{10}$).

### 3.2 Experiments and Methods

For comparison purposes, it was decided to perform two experiments using the following two datasets: (a) dataset where all missing values were removed, and (b) dataset generated by imputing missing values using linear interpolation technique mentioned in Sect. 3.1. Overview of the used datasets and their corresponding features is shown in Table 2. The total number of items in the dataset where null values were removed and the dataset where null values were imputed is 15,762 and 26,302, respectively. Input features for both datasets include hourly concentration of five air pollutants (PM$_{10}$, NO$_2$, SO$_2$, O$_3$ and CO), values of five meteorological variables
|       | Removed NaN | Linear interpolation |
|-------|-------------|----------------------|
|       | PM$_{10}$  | SO$_2$   | CO    | NO$_2$ | O$_3$ | PM$_{10}$ | SO$_2$ | CO    | NO$_2$ | O$_3$ |
| Count | 15,762      | 15,762   | 15,762| 15,762| 15.62 | 26,302     | 26,302| 26,302| 26,302| 26,302 |
| Mean  | 40.61       | 15.41    | 129.8 | 20.92  | 59.06 | 39.35      | 14.26  | 235.1 | 20.32  | 58.14  |
| Std   | 48.89       | 27.51    | 291.2 | 21.31  | 38.01 | 55.58      | 26.49  | 383   | 21.31  | 36.72  |
| Min   | 0           | 0        | 0     | 0      | 0     | 0          | 0      | 0     | 0      | 0      |
| Max   | 492         | 872      | 4598  | 206    | 435   | 756        | 872    | 5758  | 272    | 435    |
Fig. 2  Missing value statistics of the original dataset

Table 2  List of datasets and their corresponding features

| Dataset                                      | Independent variables                                      | Dependent variable |
|----------------------------------------------|------------------------------------------------------------|--------------------|
| Hourly dataset with removed null values     | PM$_{10}$, NO$_2$, SO$_2$, O$_3$ and CO                     | CAQI               |
| (15,762)                                    | Temperature, wind speed, wind direction, humidity and pressure | Date, time of the day |
| Hourly dataset with imputed null values     | PM$_{10}$, NO$_2$, SO$_2$, O$_3$ and CO                     | CAQI               |
| (26,302)                                    | Temperature, wind speed, wind direction, humidity and pressure | Date, time of the day |

(temperature, wind speed, wind direction, humidity and pressure) and two temporal variables (date and time of the day).

Within this study some of the most common machine learning methods, for the domain, were used in CAQI prediction: Support vector machine—regression (SVR) [20], Random Forest (RF) [21], XGBOOST [22], Multiple linear regression (MLR) [23] and Multilayer perceptron (MLP) [24].

3.3 Methodology

Described air pollution problem, can be categorized as a supervised learning problem, and more specifically, a regression problem. Taking into consideration a set of known input variables, the goal is to predict the output variable (i.e. target) which in this work is a continuous variable, CAQI. CAQI index is based on a scale from 1 to 100, where the lower the index, the better the air quality and higher the index, the worse the air quality. Overall index is determined based on the worst pollutant. For each
pollutant, sub-index is calculated corresponding to a grid shown in Fig. 3. The grid shows the boundaries needed to translate concentration measurement into a ranking on a scale from 1 to 100 by using linear interpolation. Highest sub-index value at a given time is going to determine the overall index. Detailed description of CAQI is provided in the work of Heich et al. [25]. Mentioned (Sect. 3.2) five machine learning methods have been used for CAQI prediction. K-fold cross-validation technique ($k = 10$) has been used to evaluate machine learning methods. All the experiments were performed using Python's scikit-learn library.

### 4 Results and Discussion

Model performance was evaluated on test dataset by comparing predictions of five tested models: SVR, RF, XGBOOST, MLR and MLP. Measuring quality of a regression model is related to how well its predictions match against the actual values. Moreover, this quality can be measured using error metrics that allow the comparison of different regressions. R-squared and RMSE are the error metrics used for measuring quality of regression in this study, and are defined as follows:

$$R^2 = \frac{\text{Variance explained by the model}}{\text{Total variance}}$$  \hspace{1cm} (1)$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$  \hspace{1cm} (2)$$

The results of all machine learning models are displayed in Table 3. Best performing model, according to regression error metrics, is Random Forest which achieved the following results: $R^2 = 0.99$ and $\text{RMSE} = 2.30$ using the dataset where...
missing values were removed, and $R^2 = 0.99$ and $\text{RMSE} = 2.58$ using the dataset where missing values were imputed. Second best performing model is XGBOOST, another ensemble method, followed by SVR, MLR and MLP. As the results show, ensemble learning methods, Random Forest and XGBOOST, outperform the conventional machine learning models. In paper published by Singh et al. [16] who also performed regression based air quality index prediction, the best performing model was also an ensemble method, i.e. Decision tree forest (DTF) with the overall performance of $R^2 = 0.96$ and $\text{RMSE} = 4.38$. Ensemble methods combine decisions of multiple models and make use of advanced ensemble techniques, e.g. bagging and boosting, which in general results in an improved overall performance compared to each individual model. Ensemble models are also known for their ability to capture both linear and non-linear relationships within the data by forming an ensemble of two or more weaker learners.

Figure 4 shows the frequency of air pollutants in generated hourly datasets. Considering air pollution in general for Bjelave neighborhood, an analysis has shown that three of the most common air pollutants affecting the quality of air are PM$_{10}$, CO and O$_3$, followed by the remaining pollutants used in the analysis. Due to the fact that CAQI is calculated based on the maximum sub-index, the exact frequency of the occurrence for each pollutant was obtained. As the results show, in the Bjelave neighborhood, CO is the most influential pollutant, which determines the CAQI value. Common sources of PM$_{10}$ include combustion activities such as motor vehicles and

### Table 3  Machine learning results

| ML model | Removed NaN $R^2$ | RMSE | Linear interpolation $R^2$ | RMSE |
|----------|-------------------|------|-----------------------------|------|
| SVR      | 0.62              | 12.81| 0.42                        | 16.23|
| RF       | 0.99              | 2.30 | 0.99                        | 2.58 |
| XGBOOST  | 0.96              | 4.19 | 0.96                        | 4.34 |
| MLR      | 0.26              | 17.89| 0.12                        | 22.56|
| MLP      | 0.53              | 14.66| 0.45                        | 15.87|

![Fig. 4 Frequency of air pollutant occurrence in hourly datasets](image-url)
industrial processes while common sources of CO include heaters or cooking equipment that runs on carbon-based fuels such as furnaces, gas ovens, gas water heaters and gas room heaters. Bjelave is an area located approximately with altitude 100 m higher than the center of the city, making it less impacted by traffic. On the other hand, it is also an area with many households that in general use furnaces for heating. It appears that this is the main reason for higher concentrations of CO then PM$_{10}$ that is measured by the Bjelave meteorological station.

Additionally, Fig. 5 illustrates the common air quality index over a three-year period. As expected, it can be noticed that CAQI is highest during the period of winter. More specifically, CAQI values start to increase shortly after the summer and reach its peak between January and February next year after which it starts to descend. One of the main reasons for having extreme episodes of PM$_{10}$ and CO only during winter time is winter inversion. During the winter, the cooler air gets trapped under the warm air which creates some sort of an atmospheric “lid”. Since the process of vertical mixing of air happens only within this layer, pollutants do not have enough of space to disperse in the atmosphere, resulting in poor air quality [26].

![CAQI over a three-year period](image)

**Fig. 5** CAQI over a three-year period
5 Conclusions

The focus of this study was prediction of air pollution using common air quality index developed by the European Union as means of standardizing air pollution measurements within European cities. Five common air quality index prediction models were used: Support vector regression (SVR), Random Forest (RF), Extreme gradient boosting (XGBOOST), Multiple linear regression (MLR) and Multilayer perceptron. Three-year period (2016–2018) data was obtained from FHMZ for Bjelave meteorological station. Data was used to generate two hourly datasets, one with removed missing values and the other with imputed missing values. Best performing model, according to regression error metrics, was Random Forest achieving the following results: $R^2 = 0.99$ and RMSE = 2.30, and $R^2 = 0.99$ and RMSE = 2.58 when missing values were removed and imputed, respectively. This study showed the potential usage of nonlinear machine learning methods in CAQI prediction. In terms of improving prediction accuracy, ensemble methods such as Random Forest appear to be the most beneficial and thus most suitable for the studied domain. Furthermore, the presented machine learning approach can also be applied to predict individual concentrations of each air pollutant. Evaluated machine learning methods were able to predict the continuous value of CAQI with high accuracy.

Acknowledgements We would like to express special thanks to Federal Hydrometeorological Institute of BiH (FHMZ) for providing access to the meteorological and air pollutant data for the Bjelave neighborhood, Sarajevo, BiH.

References

1. Saxena, P., Naik, V.: Air pollution sources, impacts and controls. CAB International (2019)
2. W. H. Organization: World Health Statistics 2018: Monitoring Health for the SDGs, Sustainable Development Goals (2018)
3. Akhtar, R., Palagiano, C.: Climate Change and Air Pollution (2018)
4. Bartelds, H., Heich, H., Léger, K.: Regional Initiative Project Common Information to European Air CAQI Air Quality Index, October 2008 (2012)
5. Gorai, A.K., Tuluri, F., Tchounwou, P.B., Ambinakudige, S.: Influence of local meteorology and NO$_2$ conditions on ground-level ozone concentrations in the eastern part of Texas, USA. Air Qual. Atmos. Heal. 8(1), 81–96 (2015)
6. Radaideh, J.A.: Effect of meteorological variables on air pollutants variation in arid climates. J. Environ. Anal. Toxicol. 07(04), 1–12 (2017)
7. Morawska, L., et al.: Applications of low-cost sensing technologies for air quality monitoring and exposure assessment: How far have they gone? Environ. Int. 116, 286–299, 01-Jul-2018
8. Kelly, F.J., Fussell, J.C.: Air pollution and public health: emerging hazards and improved understanding of risk. Environ. Geochem. Health 37(4), 631–649 (2015)
9. Z. za J. Z. FBiH: Zdravstveno stanje stanovništva i zdravstvena zaštita u Federaciji Bosne i Hercegovine 2013. Godina (2014)
10. Nethery, R.C., Wu, X., Sabath, B., Braun, D., Dominici, F.: Exposure to air pollution and COVID-19 mortality in the United States (2020)
11. Delavar, M., et al.: A novel method for improving air pollution prediction based on machine learning approaches: a case study applied to the capital city of Tehran. ISPRS Int. J. Geo-Inf. 8(2), 99 (2019)
12. Kumar, A., Goyal, P.: Forecasting of daily air quality index in Delhi. Sci. Total Environ. 409(24), 5517–5523 (2011)
13. Zickus, M., Greig, A.J., Niranjan, M.: Comparison of four machine learning methods for predicting PM10 concentrations in Helsinki, Finland. Urban Air Qual. Recent Adv. 717–729 (2011)
14. Bougoudis, I., Demertzis, K., Iliadis, L.: HISYCQL a hybrid computational intelligence system for combined machine learning: the case of air pollution modeling in Athens. Neural Comput. Appl. 27(5), 1191–1206 (2016)
15. Thompson, J.E.: Airborne particulate matter: human exposure and health effects. J. Occupat. Environ. Med. 60(5) (2018)
16. Singh, K.P., Gupta, S., Rai, P.: Identifying pollution sources and predicting urban air quality using ensemble learning methods. Atmos. Environ. 80, 426–437 (2013)
17. Amado, T.M., Dela Cruz, J.C.: Development of machine learning-based predictive models for air quality monitoring and characterization. In: TENCON 2018—2018 IEEE Region 10 Conference, vol. 2018, Oct, pp. 668–672 (2018)
18. “Sarajevo—Google maps.” Available: https://www.google.com/maps/place/Sarajevo/@43.8936377,18.2429079,11z/data=!3m1!4b1!4m5!3m4!1s0x4758cbb1ed719bd1:0x562ecda6de87b33e8m2!3d43.8562586!4d18.4130763. Accessed 27 Aug 2019
19. Noor, N.M., Yahaya, A.S., Ramli, N.A., Al Bakri Abdullah, M.M.: Filling the missing data of air pollutant concentration using single imputation methods. Appl. Mech. Mater. 754–755, 923–932 (2015)
20. Drucker, H., Burges, C.J.C., Kaufman, L., Smola, A., Vapoik, V.: Support Vector Regression Machines (1997)
21. Liaw, A., Wiener, M.: Classification and Regression by Random Forest, vol. 3, pp. 18–22, Dec 2002 (2003)
22. Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. In: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 13–17 Aug, pp. 785–794 (2016)
23. Abdullah, S., Ismail, M., Fong, S.Y.: Multiple linear regression (MLR) models for long term Pm 10 concentration forecasting during different monsoon seasons. J. Sustain. Sci. Manag. 12(1), 60–69 (2017)
24. Ramchoun, H., Amine, M., Idrissi, J., Ghanou, Y., Etaouil, M.: Multilayer perceptron: architecture optimization and training. Int. J. Interact. Multimed. Artif. Intell. 4(1), 26 (2016)
25. Van Den Elshout, S., Léger, K., Heich, H.: CAQI common air quality index—update with PM2.5 and sensitivity analysis. Sci. Total Environ. 488–489(1), 461–468 (2014)
26. TWC Team: Why do pollution levels skyrocket during winter? |The Weather Channel,” 2018. Available: https://weather.com/en-IN/india/science/news/2018-10-30-why-do-pollution-levels-skyrocket-during-winter. Accessed 14 Aug 2019