PREDICTING AND ANALYSING AIR QUALITY IN SMART CITIES USING VIRTUAL SENSORS

Gabriel Oliveira Campos[1], Gabriel Luciano Gomes[2], Felipe Domingos da Cunha[3], Leandro Aparecido Villas[4]
[1,3,4]UNICAMP, Campinas - São Paulo, Brazil, [2]PUC-MG, Belo Horizonte - Minas Gerais, Brazil

NOTE: Corresponding author: Gabriel Oliveira Campos, g265146@dac.unicamp.br

Abstract – *Virtual sensors have become essential in generating synthetic data to support smart city initiatives, mainly when physical sensors are absent or inaccessible. They play a crucial role in monitoring air pollutants, benefiting those with health sensitivities, and aiding city planning in the context of smart cities. This research focuses on creating reliable virtual sensors for air pollutant measurement through deep learning and machine learning models, providing a cost-effective solution to address the scarcity of physical sensors and data inaccuracies. The main contribution lies in improving smart cities by providing essential health-related data to mitigate the harmful effects of air pollution on residents. Continuous monitoring empowers officials and residents to reduce pollution exposure, ultimately improving public health. The study found that gradient-boosted trees are effective for daily and hourly predictions, emphasizing the vital role of virtual sensors in advancing data-driven decision-making and improving environmental quality in smart cities. The results showed that our model was better using the boosted trees method with an RMSE of approximately 10%.*

Keywords – Air pollution, Internet of Things, machine learning, urban computing, virtual sensor

1. INTRODUCTION

The rapid growth in population density in urban areas has given rise to severe consequences for human health and the environment. One of the most pressing issues is the increasing concentration of atmospheric pollutants, which compounds yearly [1]. This phenomenon is further exacerbated by the proliferation of automobiles and industrial emissions, negatively impacting air quality and contributing to the greenhouse effect [2].

Alarming statistics from the World Health Organization (WHO) estimate that approximately 4.9 million premature deaths occur annually worldwide due to air pollution [3]. In Brazil alone, air pollution was responsible for 58% of premature deaths from cardiovascular and respiratory diseases in 2016 [3]. These figures underscore the urgent need for comprehensive air quality monitoring and management. In Fig. 1, we can see that air pollution was one of the most contributing factors to the number of deaths by risk factor in the world, showing the importance of air quality in this manner.

In our ever-changing world, accurate air quality data is paramount. This data is crucial for several reasons, including protecting human health, safeguarding the environment, mitigating economic impacts, and promoting social equity. Air pollution has been linked to many health issues, including respiratory and cardiovascular diseases, cancer, and neurological disorders. It also contributes to environmental problems such as acid rain, ozone depletion, and climate change [5, 6, 7, 8, 9].

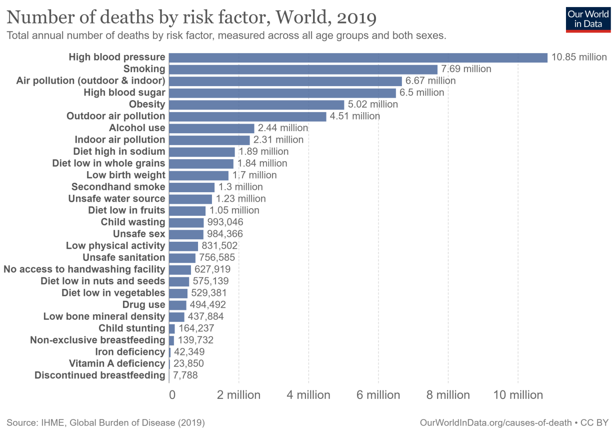

Figure 1 – Number of deaths by risk factor [4].

Although physical sensors have traditionally been used to monitor air quality, resource constraints often limit their widespread deployment. Smart cities, in particular, need help to acquire sufficient data from these sensors, leaving critical gaps in monitoring efforts. In this context, the emergence of virtual sensors powered by artificial intelligence presents a promising solution. Virtual Sensors (VS) employ mathematical models to estimate or predict physical parameters based on available data, eliminating the need for costly physical sensors and extending monitoring capabilities to previously underserved regions [10].

Our work addresses the pressing need for comprehensive air quality monitoring by creating virtual sensors for multiple harmful pollutants in various locations. We employ machine learning and deep learning techniques, evaluating their efficacy and exploring the impact of air contaminants across different cities. This research strives to answer key questions:

- What is the methodology for creating virtual sensors for air pollutants?
- Can we develop robust machine learning and deep learning models for virtual sensors?
- How do virtual sensors perform during pandemic scenarios when pollution levels vary?
- Can we forecast air quality using historical pollutant data alone, allowing real-time monitoring?

We have three main areas of contribution. To begin with, we utilize statistical techniques to delve into the functioning of virtual sensors, enabling us to grasp the connections between air pollutant levels and various influencing factors. Secondly, we construct precise virtual sensors by employing machine learning and deep learning models, striving for precise air quality assessment. These models go through extensive testing to validate their efficacy. Finally, we design virtual predictive sensors for these pollutants, allowing individuals to anticipate and prepare for upcoming changes in air quality, which is especially significant in the context of smart cities.

By addressing these critical questions and contributing to the field of air quality monitoring, we strive to advance the capabilities of smart cities and empower people to make informed decisions regarding their health and well-being. Monitoring and managing air quality is an ongoing challenge, and this work is a testament to the innovative solutions that AI-powered virtual sensors can offer to create more sustainable and healthier urban environments.

This article is organized as follows. Section 2 discusses the state-of-the-art artificial intelligence in air pollution and our contribution to this scenario; Section 3 presents the framework we proposed to use air pollution virtual sensors; Section 4 shows our research methodology using statistical analysis and an overview of the data; Section 5 offers the details of the models and the results we achieved. Finally, Section 6 summarizes our conclusion and presents future directions.

2. LITERATURE OVERVIEW

Air pollution is a significant global issue affecting millions of people's health and wellbeing. Researchers and practitioners have explored various methods to predict and forecast levels of air pollution to mitigate the impacts of poor air quality. This section gives an overview of related works in virtual sensors research, specifically on prediction and forecasting models. Section 2.1 presents studies that use regression and classification models to create machine learning methods for virtual sensors. These models aim to identify patterns and relationships between various predictors to generate accurate predictions. In Section 2.2, we discuss studies that use time series forecasting techniques. These studies typically use historical data to identify trends and patterns in the models and predict future periods. By organizing the related work in this manner, we hope to provide a comprehensive understanding of the current state of virtual sensing using machine learning, deep learning, and virtual forecasting methods.

2.1 Predicting sensor data to create virtual sensors

In [11], the authors present a method for calibrating low-cost air quality sensors using machine learning techniques. The authors demonstrate the effectiveness of their approach using real-world data collected from low-cost air quality sensors in different locations. The study results show that the virtual sensor model outperforms individual low-cost sensors in predicting air pollutant concentrations. The virtual sensor achieved a Mean Absolute Error (MAE) of 0.19 and a R^2 (R^2 or the coefficient of determination) of 0.92, while the individual sensors had an MAE of 0.38 and R^2 of 0.79. The authors also showed that their method of intelligent calibration further improves the accuracy of the virtual sensor readings, with an MAE of 0.16 and R^2 of 0.95. One limitation of the paper is that the proposed calibration and virtual sensing approach is based on a specific set of low-cost air quality sensors and may not apply directly to other sensor types or configurations.

In [12], the authors aimed to develop and evaluate machine learning models to predict the air pollution index in Indian cities. The study used data from over 20 Indian cities, including Delhi, Mumbai, and Bangalore. The study compared multiple machine learning algorithms: Support Vector Machine (SVM), XGBoost, Random Forest (RF), and K-Nearest Neighbors (k-NN). The performance of each model was evaluated using three evaluation metrics: MAE, RMSE (Root Mean Squared Error), and R^2. The results showed that the XGBoost model performed best for all cities, with an MAE value of 0.29, RMSE of 1.465, and R^2 values ranging from 0.612. The main limitation of this paper is the need to consider more carefully the heterogeneity of air pollution sources across different cities in India.

2.2 Forecasting sensor data to create virtual sensors

The authors in [13] used the Extreme Learning Machine (ELM) to forecast NO_2 and SO_2 pollutants in the following cities: Chongqing, Ganzhou, Huaihuae, and Nanchang. The effectiveness of the models generated in this work was shown through the RMSE and accuracy metrics. For the RMSE, approximately 3.949 was obtained for each analyzed model. The model proposed by the author is robust because of the more variable locations in the dataset. Still, it needs to analyze the essential variables of human health, already mentioned in the introduction, making the forecast only that reliable if there is data on the other primary pollutants. One limitation of this work is that the proposed hybrid prediction model was tested using only data from one city. It is still being determined whether the model would perform as well in other cities or regions with different air pollution characteristics. Furthermore, the paper should be compared with other state-of-the-art models for predicting air pollutant concentration, making it difficult to determine the performance of the proposed model relative to existing approaches.

In [14], the authors reviewed hybrid machine learning and forecasting models to forecast significant air pollutants. Moreover, they focused on previous work used to predict chronic airway diseases and climate change prediction. They also presented a detailed discussion about artificial intelligence methodologies. The main limitation of the reviewed results was, as they mentioned, the need for more data seasonality and the forecasting methods' dependence on the data's location.

We summarized publications on machine learning, deep learning, and forecasting methods using virtual sensors. Although our work also presents a virtual sensor prediction and forecasting, similar to others cited in this chapter, ours differs from them by using meteorological and seasonal variables to predict the concentration of atmospheric pollutants. Moreover, using these variables can prove helpful as it is widely known that both variables impact the air pollution concentration directly. Another core difference between our research and the state-of-art is the wide use of multiple cities worldwide, creating a more robust method.

3. PROPOSAL APPROACH

This section outlines our research objectives and hypotheses, which aim to use machine learning and deep learning algorithms to create virtual sensors for multiple harmful pollutants in various locations. Hence, the objectives are present in Subsection 3.1, where the virtual sensors will use data from existing physical sensors in other locations to create a prediction model and will be used in contexts where data is missing or no sensor is available. The proposal also explores the generated virtual sensors in predicting air impurities in pandemic scenarios, where accurate sensors are crucial in monitoring stations. The section also presents the research hypotheses in Subsection 3.2, which include creating virtual sensors for air pollution, predicting pollutant concentration in different cities, the models' robustness to external data, and the model's ability to predict pollutants in a pandemic scenario. The research aims to contribute to the existing knowledge on creating virtual sensors for air pollutants and improving air monitoring in smart cities.

3.1 Research objectives

This article comprehensively explores machine learning and deep learning techniques to establish virtual sensors for monitoring harmful pollutants across diverse geographical locations. Our core objective is to understand the dynamics of air contaminants within urban centers, focusing on the pivotal role played by various model variables. Our framework involves creating virtual sensors through machine learning and deep learning algorithms, utilizing data from existing physical sensors in various locations to develop predictive models. A central research goal is to apply these virtual sensors in scenarios characterized by data gaps in the target variables or a lack of physical sensors for these pollutants. Additionally, we investigate the utility of these virtual sensors in predicting air impurities during pandemic scenarios, recognizing the heightened demand for precise sensor data in smart cities' monitoring stations.

Our study involves meticulously evaluating diverse techniques within each learning paradigm to identify the most effective machine learning or deep learning method for virtual sensor creation. We consider notable deep learning models, including artificial neural networks, convolutional neural networks, deep learning belief networks, and transfer learning, which have demonstrated excellence in regression tasks across various applications. Furthermore, we consider the remarkable performance of ensemble methods in specific regression applications. Consequently, a careful assessment of various methodologies becomes imperative to identify the optimal learning model within the context of air pollution prediction.

The implementation of virtual sensors remains contingent upon a substantial input of data from physical sensors. Our study suggests that the model trained solely on physical sensor data to predict a single pollutant necessitates the presence of a comprehensive monitoring station to enable predictions via virtual sensors. This requirement arises due to the inclusion of additional variables utilized during training, such as temperature and humidity, which contribute to the complexity of the predictive process. Therefore, while virtual sensors offer promise in environmental monitoring, their efficacy hinges upon the availability of comprehensive data in-

puts from physical sensors and consideration of relevant environmental variables.
A significant outcome of our model is the development of real-time virtual sensors capable of forecasting pollutant concentrations. These sensors provide essential information on current and upcoming air quality conditions, catering to individuals with heightened health sensitivities. Moreover, we introduce the creation of a virtual sensor through a forecasting approach, relying exclusively on historical data for a designated pollutant. This simulation addresses scenarios where all data from a monitoring station is irretrievably lost, leaving only historical pollutant data for analysis. Additionally, we aspire to predict air quality by leveraging historical data related to a single pollutant.

We further encompass the formulation of hypotheses that guide our objectives. These hypotheses are guiding principles for analyzing and interpreting the data collected. We investigate these hypotheses by employing rigorous statistical analysis methods, subsequently using the results to support or refute their respective premises. Our hypotheses span various aspects of the research, including the prediction of virtual sensor accuracy, the evaluation of meteorological variables' impact on the model, and an analysis of the influence of the COVID-19 pandemic on the predictive capabilities of virtual sensors. Through the formulation and testing of hypotheses, our proposal aims to contribute to the existing body of knowledge concerning the creation of virtual sensors for air pollutants and the enhancement of air quality monitoring within smart cities.

3.2 Research hypotheses

To advance the development of virtual sensors for air pollution prediction, we have consolidated the hypotheses into a comprehensive framework that aligns with our objectives. This framework is founded on the following key assumptions and objectives:

1. **Hypothesis 1:** The potential to create virtual sensors for air pollution prediction exists through machine learning and deep learning models. As the demand for proficient prediction techniques in the Internet of Things (IoT) domain continues to escalate, the inadequacy of conventional statistical methods in the face of extensive and intricate data generated by IoT devices becomes evident. In contrast, machine learning and deep learning models demonstrate their aptitude for processing complex, high-dimensional data, often yielding more accurate and resilient predictions. We illustrate the viability of this hypothesis through the development of the "São Paulo Hourly VS Model," demonstrating the creation of an air pollution virtual sensor.

2. **Hypothesis 2:** The concentration of air pollutants in one city can be harnessed to predict those in other cities, fostering the development of the "Specific Cities Daily VS Model." Following the generation of the initial virtual sensor at an hourly granularity, this progression aims to create a daily granularity sensor for specific pollutants. We employ data from multiple urban centers worldwide instead of relying solely on São Paulo city data. This approach explores the potential of using data from diverse urban areas to enhance the learning model for virtual sensors, accommodating different data granularities.

3. **Hypothesis 3:** The virtual sensor models exhibit robustness when confronted with external data, thus validating the "Generic Cities Daily VS Model." This phase of the research involves the creation of multiple virtual sensors across numerous cities, explicitly assessing the robustness of employing data from distinct urban centers to generate virtual sensors for one another. By evaluating model precision and accuracy with a testing city and others as the training dataset, we aim to establish the model's robustness to variations in data sources and regions.

4. **Hypothesis 4:** The virtual sensor models remain effective in a pandemic scenario, as examined in the "Pandemic São Paulo VS Model." A significant challenge arises when external factors dramatically alter the environmental conditions of a city, as exemplified by the COVID-19 pandemic. This hypothesis explores the adaptability and continued effectiveness of virtual sensor models during pandemic-induced changes in air quality, even when the urban landscape experiences considerable transformation.

5. **Hypothesis 5:** The ability to create a virtual sensor is not contingent on monitoring station data availability, as demonstrated in the "São Paulo Forecasting Sensor." In scenarios where data from monitoring stations is absent, we have devised a model that utilizes forecasting techniques, leveraging historical pollutant values to generate accurate data. This model accounts for seasonality, aiming to create virtual sensors through forecasting methods.

Our approach introduces an innovative methodology for predicting carbon monoxide and particulate matter in the air, incorporating meteorological variables to address weather conditions' impact on air pollutant concentrations. By conducting this comprehensive research, we aim to provide a robust and precise virtual sensor model applicable across diverse geographic locations, thereby overcoming the limitations inherent in current approaches.

4. METHODOLOGY

This section provides a comprehensive overview of the data collection, preprocessing steps, and statistical analysis techniques used to develop virtual sensors. The section begins in Subsection 4.1 by describing the data-gathering process and highlighting the sources from which data was collected. subsection 4.2 details the steps in preprocessing the data to ensure its accuracy and reliability. This section also discusses the various methods utilized in developing virtual sensors and their corresponding parameters. The parameters were carefully selected based on their ability to predict the targeted variables accurately.

4.1 Dataset acquisition

In pursuing pollutant data, we delved into various open datasets to access pertinent information. Amidst our extensive search, we uncovered a comprehensive global repository known as the World Air Quality Project (AQICN), which diligently aggregates official data concerning air pollution levels from government websites. This invaluable resource offers unrestricted access to data from cities worldwide, and served as the foundation for our study [15].

The pollutant data sourced from AQICN underwent a substantial transformation, meticulously converted into the United States Air Quality Index (US AQI) format. The US AQI represents a standardized methodology for pollutant analysis, skillfully amalgamating concentration levels of multiple pollutants into a singular, dimensionless value. This approach facilitates the comparison of pollutant levels with legally defined thresholds for various pollutants. We also employed a web crawler [16] to extract meteorological data from the CustomWeather API [17] about the diverse cities featured in our study. Furthermore, we incorporated information about the seasonal impact on air quality, ensuring that seasonal variations were duly considered in the analysis.

The culmination of our data collection efforts is paramount in developing our models and the validation process for each use case. To ensure the robustness of our models, we employed two distinct techniques for validation: blind randomization of training and testing data and cross-validation across all cities, where one city served as the exclusive testing dataset. We expound upon the intricacies of each method in the forthcoming results section.

4.2 Models and hyperparameters

Fine-tuning and grid search are essential steps in optimizing machine learning models. Fine-tuning involves adjusting the hyperparameters of a pre-trained model to improve its performance on a specific task [18]. Fine-tuning enables us to adjust these parameters to perform better on a specific task.

Grid search is a technique for hyperparameter tuning that involves searching for the optimal combination of parameters by evaluating the model's performance on a grid [19]. This process consists of specifying a range of values for each hyperparameter and assessing the model's performance on all possible combinations of these values. By systematically evaluating each variety of parameter, grid search can identify the optimal mix that results in the best performance on the task.

Both fine-tuning and grid search are essential in improving machine learning models' performance. With fine-tuning, the model may perform optimally on the specific task for which it was trained. To achieve the best performance, we can identify the optimal combination of hyperparameters with grid search. Therefore, these techniques are critical in ensuring that machine learning models can effectively apply to real-world problems. Additional information about the processing time, best hyperparameters for each model, and performance metrics are available in [20].

5. RESULTS

In this section, we will present the results of our study, which focused on testing various hypotheses related to predicting air pollutant concentrations using virtual sensors and artificial intelligence models. In subsections 5.1, 5.2, 5.3, 5.4, and 5.5, we show the results of our proposed hypotheses. We will also discuss the Shapley values metric in Subsection 5.6, used to determine the feature importance of the artificial intelligence models. This metric helps us identify the most significant variables for predicting each pollutant. We will discuss how it contributes to understanding the relationships between pollutants and meteorological variables. Overall, this section will provide a comprehensive overview of our study's findings and contribute to the broader field of air quality research.

5.1 Hypothesis 1: São Paulo hourly regression model

In this subsection, we unveil the outcomes derived from utilizing hourly data sourced from the Qualar system, an archive of historical air quality information made available by the São Paulo government. We developed a virtual sensing model harnessing this hourly dataset, establishing the foundational model for our research endeavors. The principal objective underlying the creation of this model was to validate the viability of crafting a virtual sensor for air pollution through the deployment of hourly data.

After creating the base model for hourly data, we created a simple virtual sensor. Hence, we utilized Machine Learning (ML) and Deep Learning (DL) models to identify the most effective models for predicting a single city's air quality with hourly granularity. Figures 2 and 3 present the RMSE for each pollutant and the scatter plot of these results for the boosted trees model, which had the lowest RMSE. We also note that Fig. 2 only displays the best-performing ML and DL models, and Fig. 3 only shows the best model, the boosted trees. This approach allowed us to identify the most effective models for predicting air quality with hourly granularity, providing a valuable tool for air quality management and monitoring.

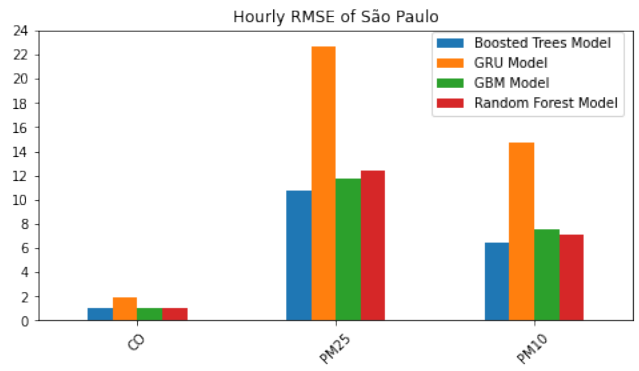

Figure 2 – São Paulo hourly regression model RMSE.

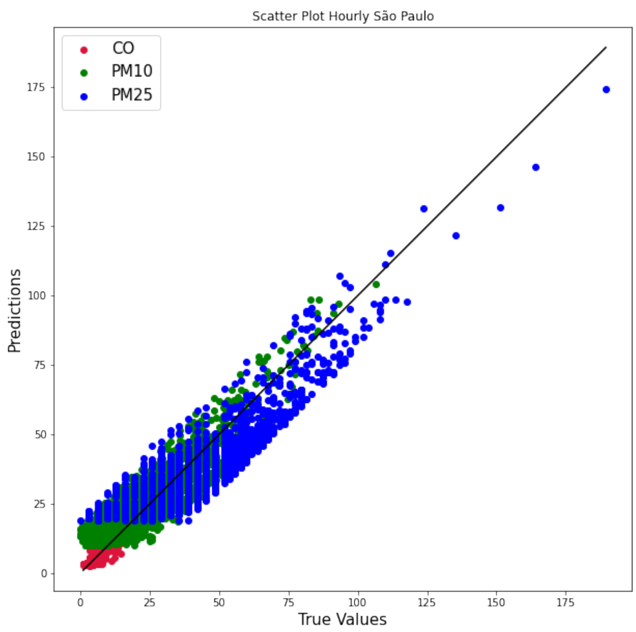

Figure 3 – São Paulo hourly regression boosted trees scatter plot.

Our analysis validated the first hypothesis as the virtual sensor generated consistently produced accurate results, as shown in the scatter plot. Each data point in the scatter plot represents testing data, and the closer it is to the central line, the more accurate the model is. We observed that the variation of the results is low, and there are no significant outliers. Although the air pollution concentration can vary up to 175, the root mean squared error for the boosted trees model was low, indicating good performance with RMSE values of 1.11, 10.91, and 6.7 for the pollutants CO, $PM_{2.5}$ and PM_{10}, respectively. Therefore, we can confidently conclude that the virtual sensor produced reliable and accurate results in this first hypothesis.

5.2 Hyphotesis 2: Specific cities daily regression model

In the second hypothesis, we investigated further by dividing the dataset into training and testing data. We used all cities in the training data and the last 180 days of a specific town for testing. This allowed us to assess the performance of the virtual sensing models in different cities. Building on the success of the previous hypothesis, we aimed to create a better model for São Paulo City and other major cities worldwide. However, to achieve consistency in the data granularity, we used daily data instead of hourly data, as the former was more readily available.

In this model, our objective is to evaluate the performance of various ML and DL models to determine if the boosted trees method remains the most effective in producing accurate and precise results. To validate this hypothesis, we aim to obtain results similar to our previous analysis, confirming that the virtual sensor model has a low RMSE, high accuracy and precision, and minimal variation without significant outliers. By achieving these benchmarks, we can confirm that the virtual sensor model is reliable and consistent, providing accurate air pollution predictions for various cities worldwide.

Due to the increased scope of our analysis, which involves multiple cities and a significant amount of data, we created separate plots for each pollutant. We present the scatter plot and the RMSE plot for each pollutant, providing a detailed analysis of the results. In figures 4, 5, 6, 7, 8 and 9, we can observe these plots for each of the pollutants. This approach allows us to understand better each model's performance in predicting different pollutants in various cities.

After analyzing the second hypothesis, we have confirmed that it holds. The results from the virtual sensor using daily data were even better than the previous model with hourly granularity. For each city, there were no outliers detected in the scatter plot, and the mean RMSE across all cities was approximately 5.12, 5.92,

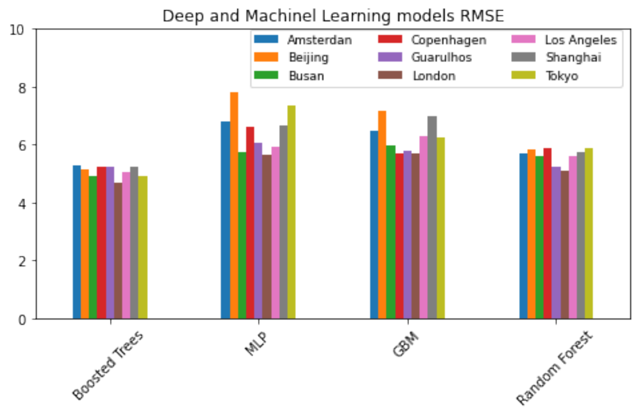

Figure 4 – Multiple cities daily CO model RMSE.

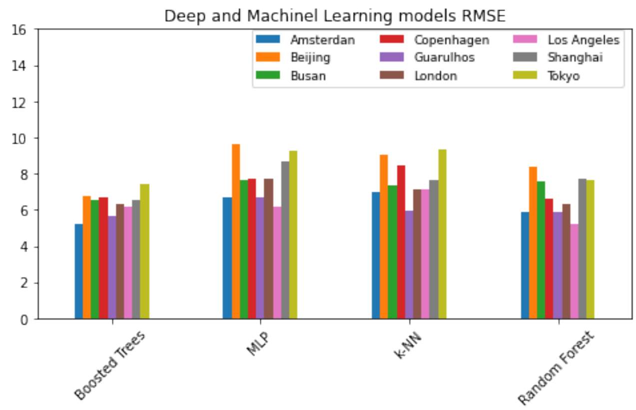

Figure 6 – Multiple cities daily PM_{10} model RMSE.

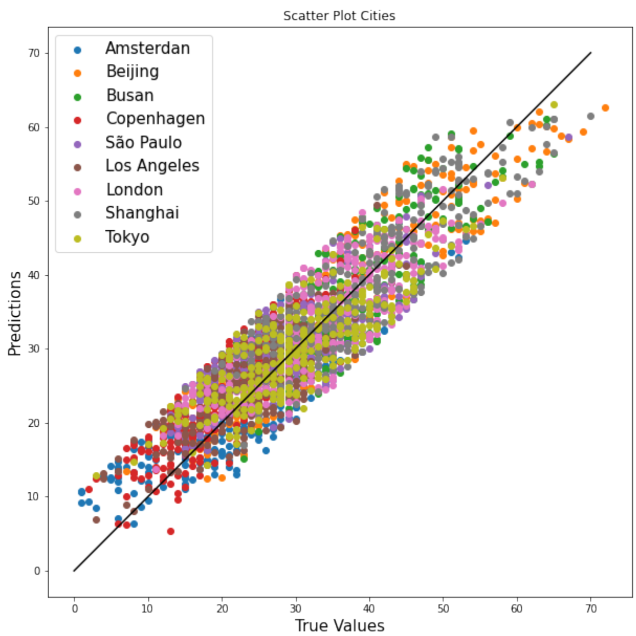

Figure 5 – Multiple cities daily CO model scatter plot.

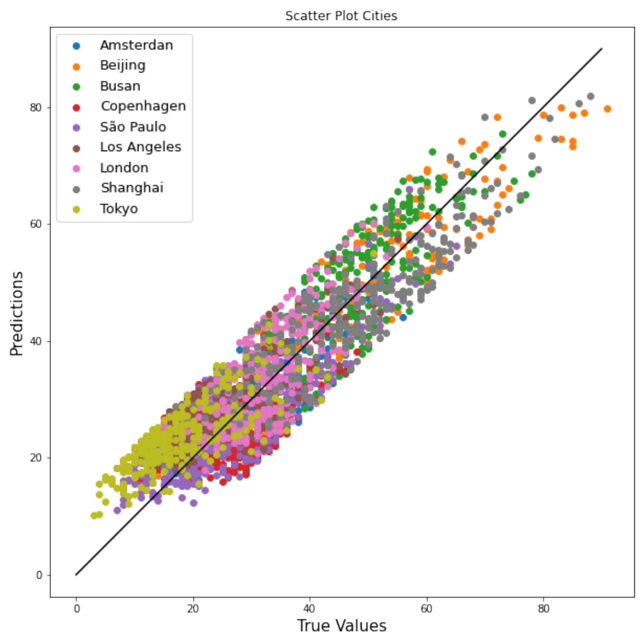

Figure 7 – Multiple cities daily PM_{10} model scatter plot.

and 7.98 for CO, PM_{10}, and $PM_{2.5}$, respectively. Once again, the Boosted Trees model produced the most accurate results compared to other models, and there was lower variation between the cities. This result demonstrates the robustness of using data from other towns to predict air pollution concentrations daily. Additionally, since the results for the hourly and daily models for São Paulo were similar, it is possible to assume that the model would exhibit similar robustness if sufficient hourly data were available for other cities.

5.3 Hypothesis 3: Generic daily regression model

In the third hypothesis, we aimed to further test the robustness of the virtual sensor model by randomizing both the training and testing data. Unlike the previous theory, where we used specific data for training and testing, we aimed to test whether the model could still produce accurate results regardless of the data set. By randomizing the data, we tested the model's ability to predict air pollution concentrations in any given city with any given collection of data. Suppose the model can still generate accurate results with random data. In that case, the dataset is robust, and the model can predict air pollution concentrations in various situations. This hypothesis is the final step in testing the effectiveness of the virtual sensor in predicting air pollution in typical cases.

As previously discussed, the performance of the models will be evaluated using the RMSE metric, which measures the average distance between predicted and actual values. The results are presented in figures 10 and 11, which portray the best models for each pollutant. It is evident that the boosted trees method consistently outperforms the other models for all air contaminants in terms of RMSE. This behavior suggests that the boosted trees algorithm is more accurate and reliable for predicting air pollution concentration.

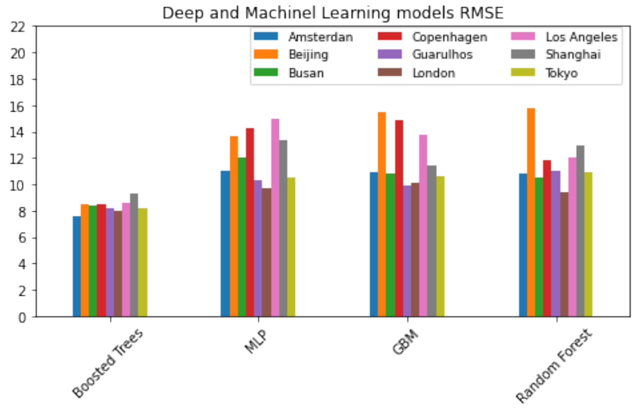

Figure 8 – Multiple cities daily $PM_{2.5}$ model RMSE.

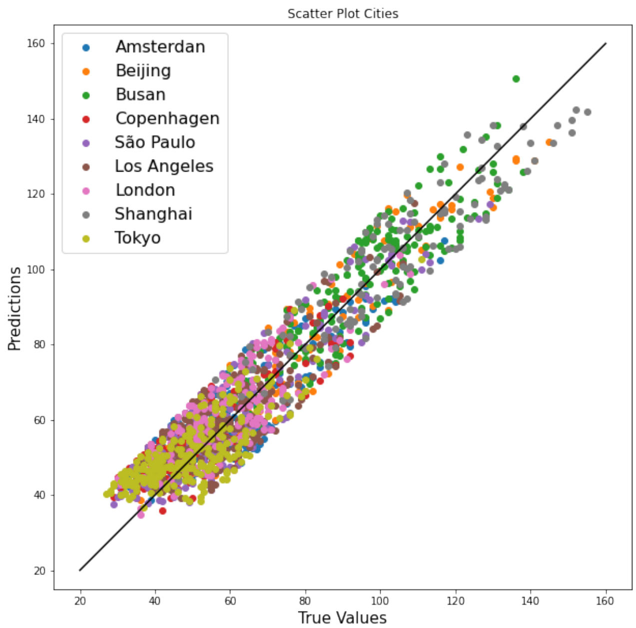

Figure 9 – Multiple cities daily $PM_{2.5}$ model scatter plot.

In this hypothesis, we confirmed the robustness of the virtual sensor by randomly selecting training and testing data from multiple cities for each pollutant. We were able to confirm this hypothesis as the results were consistently close to the specific cities models, with no signs of outliers, low variation, and a good RMSE. These findings further support the notion that virtual sensors can provide accurate and reliable data for air pollution monitoring.

After analyzing the third hypothesis, we demonstrated the reliability of using virtual sensors in air pollution monitoring. We created a variety of virtual sensors for both hourly and daily granularity. In all of them, we continued to seek more evidence that proves the reliability of using virtual sensors in air pollution.

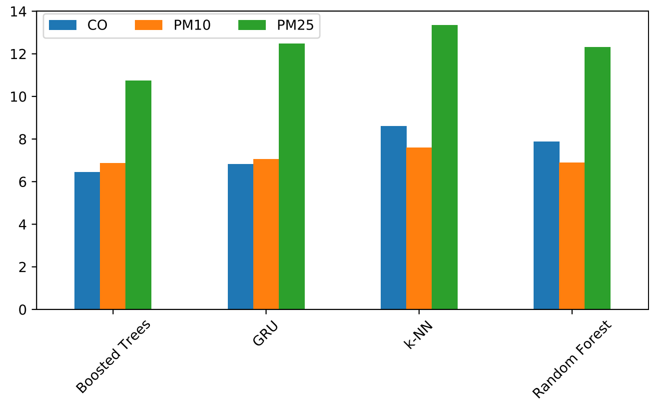

Figure 10 – Generic daily regression model RMSE.

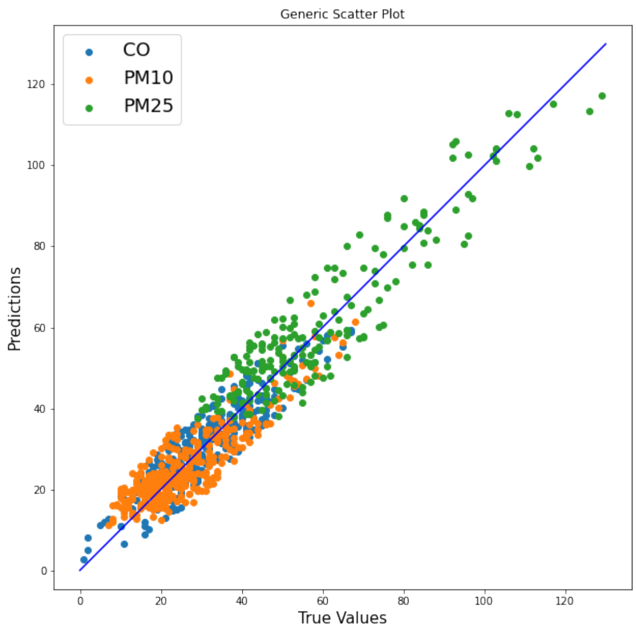

Figure 11 – Generic daily regression model scatter plot.

5.4 Hypothesis 4: Pandemic São Paulo hourly regression model

As the COVID-19 pandemic continues to affect the world, various measures have been implemented to mitigate its spread. One of the most significant measures is the lockdown, where people must stay inside their homes and only leave for essential purposes. This decision has significantly reduced industrial activities and transportation, leading to a noticeable improvement in air quality in many cities worldwide. In this hypothesis, we tested the effectiveness of using air pollution virtual sensors in a changing context inside towns. The aim was to determine if the virtual sensors could accurately predict air pollution concentrations during the lockdown, which saw a significant change in the air quality of most cities.

In this fourth hypothesis, we aimed to test the usage of air pollution virtual sensors in a changed context within cities during the COVID-19 pandemic. We focused on the hourly granularity data of São Paulo city as it provided a real-world scenario with a lower data granularity. However, we believe the results obtained can be generalized to other cities, given the robustness of the virtual sensor replication. We used the model trained in the first hypothesis to test the fourth hypothesis, which included only the São Paulo hourly data. Then, we applied this same model to March and July, the peak months of the COVID-19 lockdown in Brazil.

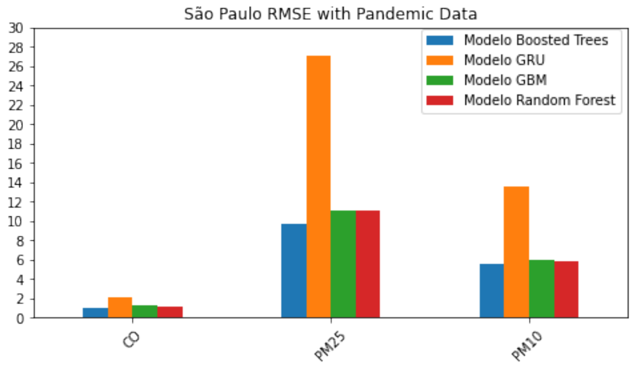

Figure 12 – Pandemic hourly regression model RMSE.

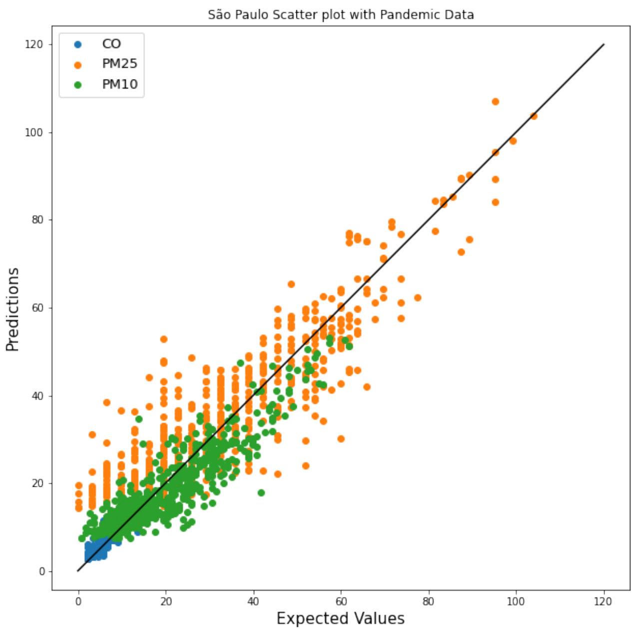

Figure 13 – Pandemic hourly regression model scatter plot.

After analyzing the fourth hypothesis, we observed, in figures 12 and 13, that the virtual sensor results were still reliable even with a change in context inside the cities, specifically during the COVID-19 lockdown when air quality improved due to decreased traffic and industrial activities. The RMSE values were 1.12, 5.96, and 10.03 for CO, PM_{10}, and $PM_{2.5}$, respectively. Some outliers were observed, especially for $PM_{2.5}$, but the RMSE was still as good as in the first theory for the boosted trees model. Despite the presence of outliers, the hypothesis was confirmed due to the good RMSE and low variation of PM_{10} and CO. Additionally, the model can retroactively learn, meaning that the actual lockdown data could be used to remove outliers and produce good results with just a few lockdown data points.

5.5 Hypothesis 5: São Paulo hourly forecasting model

In the fifth hypothesis, we addressed the issue of network loss in air pollution stations. When the entire network goes down, it becomes impossible for the virtual sensor to produce any results. To address this issue, we proposed a virtual forecasting sensor that only relies on historical data of the desired pollutant. However, the accuracy and precision of such a sensor heavily depend on the presence of seasonality in the data. To test the forecasting model of the hourly granularity data of São Paulo, we employed the Facebook Prophet model, a tool for time series forecasting. The goal was to determine if a significant seasonality was present in the data that could be used to generate more accurate and precise predictions.

Figures 14, 15, and 16 display the results of applying the Facebook Prophet model to the hourly granularity data of São Paulo to verify the presence of seasonality in the data. The figures depict the predicted values of each data point in the dataset and future predictions of the pollutant. The Prophet model uses a cross-validation method where 10% of the data is used for testing and repeatedly changed until all the data has been used as a test once. Each figure shows a set of black dots representing the real data and blue dots representing the predicted values obtained using the forecasting model. It is important to note that the lack of seasonality in the data resulted in the model being unable to predict values that deviated significantly from the mean values observed in the historical data. The model relies solely on past observations and is not designed to predict rare or uncommon events, such as air pollutants with high concentrations. Therefore, these events are not part of the historical data and are difficult for the model to capture. As a result, the model is more suited for predicting values within the range of historical data and may not perform as well when attempting to predict extreme or unusual events.

The virtual forecasting sensor was proposed in the fifth hypothesis to address the problem of network loss, which makes it unsuitable for virtual sensors to produce any results. This sensor would use only historical data of the wanted pollutant to make predictions. However, the results of the Facebook Prophet model showed a lack of seasonality in the data, making it challenging to create an accurate forecasting model. This lack of seasonality was a significant problem because forecasting models require significant seasonality to produce more precise and accurate data. Therefore, we refuted the final hypothe-

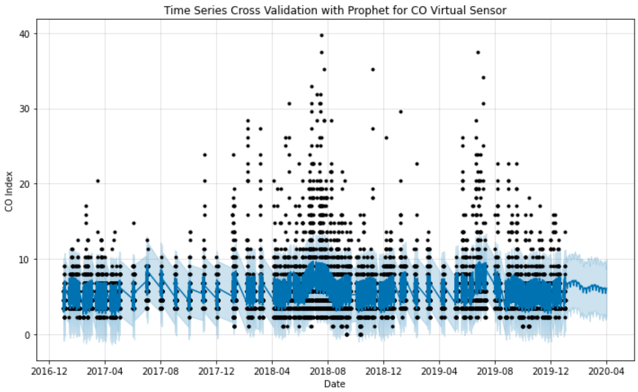

Figure 14 – Carbon monoxide forecasting using Prophet.

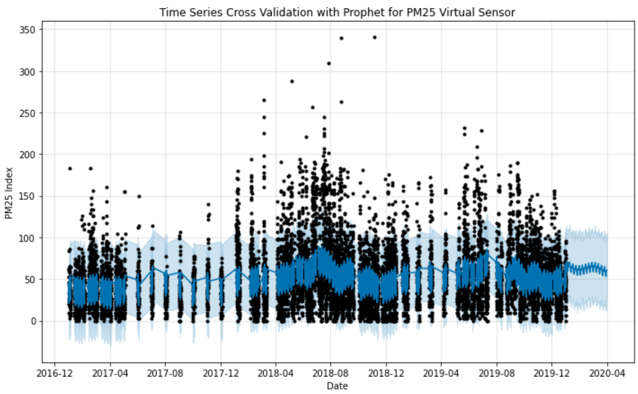

Figure 15 – $PM_{2.5}$ forecasting using Prophet.

sis due to this problem. However, if there were a more restrictive granularity, such as every minute, it could generate better results than the ones shown in this subsection.

5.6 Most important variables with Shapley values

To evaluate the most important features for creating the air pollutant virtual sensor, we used Shapley values. This statistical method determines the importance of each variable in predicting the ground truth. This method allowed us to create a more explainable AI model, which is an essential factor in environmental pollution prediction, where stakeholders need to understand the contributing factors and sources of pollution to take practical actions to mitigate it.

To apply the Shapley values method, we used the boosted trees model trained on the hourly granularity data of São Paulo city. The results of the most important features for CO are presented in Fig. 17. We tested with the pollutants PM25 and PM10, but since they showed the same results as CO, we opted not to show the plots in the results. The magnitude of these values reflects the relative importance of each variable in the model's prediction. The magnitude of each variable in Shapley values depends on how much it contributes to

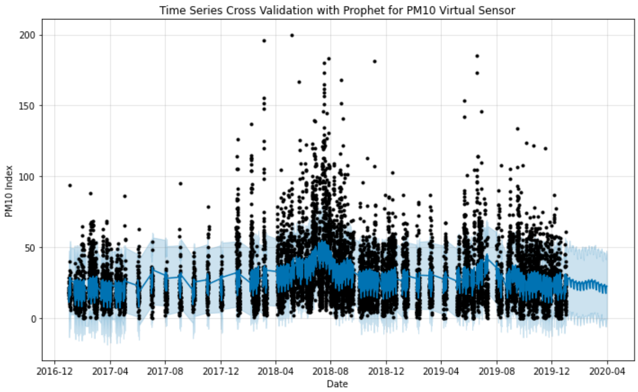

Figure 16 – PM_{10} forecasting using Prophet.

the model's prediction for a given instance. The contribution is measured by comparing the model's output with and without the variable's value for that instance. The difference in the model's production is attributed to that variable's contribution, and the Shapley value is the average of all such differences across all possible permutations of variables. The feature importance can be positive or negative, indicating whether its value increases or decreases the model's prediction for that instance.

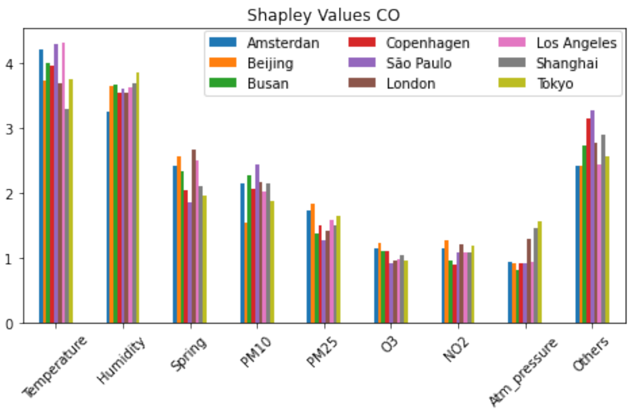

Figure 17 – Carbon monoxide Shapley values in specific cities daily regression model.

The analysis of the Shapley values showed that for carbon monoxide virtual sensors, meteorological variables were the principal features that impacted the model's correct prediction. This result signifies that these variables must be considered in the virtual sensor creation process. As for PM_{10} and $PM_{2.5}$ virtual sensors, the most crucial variable was found to be the pollutants themselves, indicating the presence of a strong correlation between them. Other air pollutants such as SO_2 and NO_2 were also found to be essential features for these virtual sensors.

In summary, the Shapley values analysis provided insights into the most important variables for each virtual pollutant sensor, which can be used to improve the accuracy and performance of the models.

6. CONCLUSION AND FUTURE DIRECTIONS

This study presented virtual sensing models' effectiveness in monitoring carbon monoxide air pollutants in urban environments. The research demonstrated the capability of these models to supplement existing monitoring networks, offering a cost-effective and practical method to analyze various pollutants. Our findings revealed the similarity in pollutant behaviors across diverse urban centers, indicating the transferability of virtual sensing models to unexplored areas without significant loss of accuracy.

The assessment of virtual sensor model performance highlighted the superiority of the boosted trees model in accuracy and processing time. However, the study also revealed limitations in forecasting accuracy due to the variability in environmental factors and the absence of seasonality in historical data.

Overall, this work confirmed the viability of virtual sensors as reliable, accurate, and precise tools for measuring air pollutants in urban settings, providing crucial support for cities lacking extensive monitoring infrastructure. These virtual sensors have the potential to contribute valuable data for Smart Health Communities (SHC), aiding in monitoring air pollution's impact on public health and guiding appropriate interventions.

Future research should focus on developing virtual sensor models that can operate effectively even with missing or incomplete data. Understanding the critical variables for accurate predictions when physical sensors become inoperative is essential. Investigating the creation of virtual sensors in scenarios with incomplete data is crucial for practical applicability and efficiency.

Additionally, exploring the integration of deep learning techniques, like deep belief networks, holds promise in enhancing the precision and resilience of virtual sensor models. These advanced models have the potential to capture more intricate relationships between pollutants and environmental factors, leading to more accurate predictions and a more comprehensive understanding of air quality. Such advancements can significantly contribute to improving environmental monitoring and the development of smarter and healthier cities.

REFERENCES

[1] Tony Moran. *Poluição por incêndios florestais no Brasil agrava qualidade do ar em cidades distantes.* [Online; accessed 17-December-2020]. 2020.

[2] Nicole Zulauf, Janis Dröge, Doris Klingelhöfer, Markus Braun, Gerhard M Oremek, and David A Groneberg. "Indoor air pollution in cars: an update on novel insights". In: *International Journal of Environmental Research and Public Health* 16.13 (2019), p. 2441.

[3] Amanda Mendes. *Mortes devido a poluição.* [Online; accessed 13-July-2020]. 2019.

[4] Institute of Health Metrics and Evaluation. "Deaths by risk factor". In: *Our World in Data* (2019).

[5] Lilian LN Guarieiro, Pérola C Vasconcellos, and Maria Cristina Solci. "Poluentes atmosféricos provenientes da queima de combustíveis fósseis e biocombustíveis: uma breve revisão". In: *Revista Virtual de Química* 3.5 (2011), pp. 434–445.

[6] Camila Trolez Amâncio and Luiz Fernando Costa Nascimento. "Asma e poluentes ambientais: um estudo de séries temporais". In: *Revista da Associação Médica Brasileira* 58.3 (2012), pp. 302–307.

[7] Magali Vieira Santiago Bucco. "Construção e testes de validação de amostradores passivos para dióxido de nitrogênio e ozônio". In: *Dissertação* (2010).

[8] F de F Peres. "Meio Ambiente e Saúde: os efeitos fisiológicos da poluição do ar no desempenho físico-o caso do monóxido de carbono (CO)". In: *Arquivos em movimento* 1.1 (2005), pp. 55–63.

[9] S. Twomey. "The Influence of Pollution on the Shortwave Albedo of Clouds". In: *Journal of the Atmospheric Sciences* 34.7 (July 1977), pp. 1149–1152. ISSN: 0022-4928.

[10] Sanem Kabadayi, Adam Pridgen, and Christine Julien. "Virtual sensors: Abstracting data from physical sensors". In: *2006 International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM'06)*. IEEE. 2006, 6–pp. DOI: 10.1109/WOWMOM.2006.115.

[11] Martha Arbayani Zaidan, Naser Hossein Motlagh, Pak L Fung, David Lu, Hilkka Timonen, Joel Kuula, Jarkko V Niemi, Sasu Tarkoma, Tuukka Petäjä, Markku Kulmala, et al. "Intelligent calibration and virtual sensing for integrated low-cost air quality sensors". In: *IEEE Sensors Journal* 20.22 (2020), pp. 13638–13652. DOI: 10.1109/JSEN.2020.3010316.

[12] K Kumar and BP Pande. "Air pollution prediction with machine learning: A case study of Indian cities". In: *International Journal of Environmental Science and Technology* (2022), pp. 1–16. DOI: https://doi.org/10.1007/s13762-022-04241-5.

[13] Wei Sun and Chenchen Huang. "A hybrid air pollutant concentration prediction model combining secondary decomposition and sequence reconstruction". In: *Environmental Pollution* 266 (2020), p. 115216. ISSN: 0269-7491. DOI: https://doi.org/10.1016/j.envpol.2020.115216.

[14] Shankar Subramaniam, Naveenkumar Raju, Abbas Ganesan, Nithyaprakash Rajavel, Maheswari Chenniappan, Chander Prakash, Alokesh Pramanik, Animesh Kumar Basak, and Saurav Dixit. "Artificial Intelligence Technologies for Forecasting Air Pollution and Human Health: A Narrative Review". In: *Sustainability* 14.16 (2022), p. 9951. DOI: https://doi.org/10.3390/su14169951.

[15] Muthukumar Kumar. *Real time air pollution map of the world: The World Air Quality Index project.* 2015.

[16] Md Abu Kausar, VS Dhaka, and Sanjeev Kumar Singh. "Web crawler: a review". In: *International Journal of Computer Applications* 63.2 (2013).

[17] Xiang Wan, Wenqian Wang, Jiming Liu, and Tiejun Tong. "Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range". In: *BMC medical research methodology* 14.1 (2014), pp. 1–13. DOI: https://doi.org/10.1186/1471-2288-14-135.

[18] Lisha Li, Kevin G Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. "Hyper-parameter tuning in deep learning". In: *arXiv preprint arXiv:1810.05934* (2018).

[19] James Bergstra and Yoshua Bengio. "Random search for hyper-parameter optimization". In: *Journal of Machine Learning Research* 13.Feb (2012), pp. 281–305.

[20] Gabriel Oliveira Campos, Leandro Aparecido Villas, and Felipe Domingos Da Cunha. "Analysis of Pandemic Atmosphere Pollution Data Using Virtual Sensors in São Paulo City". In: *2022 18th International Conference on Distributed Computing in Sensor Systems (DCOSS)*. IEEE. 2022, pp. 314–321. DOI: 10.1109/DCOSS54816.2022.00058.

AUTHORS

Gabriel Oliveira Campos received his bachelor's degree in computer science in 2019 from the Pontifical Catholic University of Minas Gerais and master's degree in computer science in 2023 from the University of Campinas, Brazil. He is currently a senior data scientist in Catho LTDA in São Paulo, Brazil. His interests include data science, the Internet of Things, urban computing, artificial intelligence and virtual sensors

Gabriel Luciano Gomes received his bachelor's degree in computer science in 2019 from the Pontifical Catholic University of Minas Gerais and is a master's student in computer science at the University of Campinas, Brazil. He is currently working as a research and development engineer at Synopsys France. His interests are machine learning, federated learning, and differential privacy to improve urban mobility.

Leandro Aparecido Villas finished his doctoral degree in computer science at the Federal University of Minas, Brazil, in 2012. He had the doctoral thesis selected among the top six in the Brazilian Computer Society. He was awarded the Excellence Research Award from PARADISE Laboratory - University of Ottawa. Currently, he is a CNPq researcher level 1D, associate professor of computer science at the Institute of Computing at the University of Campinas, and coordinator of the Hub of Artificial Intelligence and Cognitive Architectures at Unicamp (https://hiaac.unicamp.br/en/). Dr. Villas has already published 80+ papers in international journals and 160+ in conferences. Nine of those papers received the best paper award in IEEE/ACM and SBC conferences. He was recently named among the top 1% of scientists with the most significant impact in the world, according to a study by Stanford University/Plos Biology.

Felipe Domingos da Cunha's background includes a bachelor's degree in computer science obtained in 2006 from the Pontifical Catholic University of Minas Gerais. He has a master's degree (2009) and Ph.D. in computer science (2016) from the Federal University of Minas Gerais, where he also completed a research period at the Centre de Recherche Inria Saclay in France. He is a professor and heads the computer science department at PUC Minas. His professional expertise lies in computer science, focusing on computer networks. His primary research areas encompass wireless sensor networks, mobile computing, ubiquitous computing, the Internet of Things, and urban computing.

