Air Pollution Forecasting with Novel Machine Learning Approach and Public Opinion Analysis in Valencia

Nam Le and Sebastian Swobada†

Emails: nle@etsinf.upv.es; sswobod@etsinf.upv.es;
†These authors contributed equally to this work.

Abstract

In recent years, air pollution has become increasingly serious in various cities worldwide, including Valencia. Consequently, accurately predicting the level of pollution plays a crucial role in finding effective solutions to combat this issue. This research paper focuses on utilizing the LSTM model to forecast concentrations of two main pollutants (specifically NO2 and PM10) and examining the opinions of Valencia residents on a new solution for air pollution through a short survey. The LSTM model proves to be efficient in predicting pollutant levels and identifying areas at risk of excessive pollution, and the survey highlights that residents have a moderate level of satisfaction with the current air quality and express significant support for the proposed air sensor solution. The findings underscore the importance of addressing air quality challenges by implementing measures such as expanding green spaces and enhancing urban traffic management.

Keywords: Valencia, Pollution, Air sensor, Survey Analysis, LSTM model

1 Introduction

Valencia is voted the best city for expats to live in the world in 2022 voted by Internations - the world's largest expat community [1]. It is famously known for the big event Fallas organized every March with a lot of activities such as mascletas and fireworks. However, in recent years, with the increase in the number of people living in Valencia, combined with the excessive amount

of fireworks and smoke from burning ninots, air pollution gradually becomes one of the main problems in Valencia. Despite the Covid-19 pandemic when lockdowns were applied, the air quality has not been significantly improved, and the O3 pollution level was even found to increase (Briz-Redón et al., 2020) [2]. At the moment, Valencia's air pollution level is low, but it is noted that there is an implicit threat of air pollution, and the consequences will turn out to be more severe than ever. This paper aims to predict the pollution level of Valencia in the future, with a focus on potential areas that can be polluted soon. This level can be measured by calculating the concentration of different pollutants in the air, mainly nitrogen dioxide, particle matter (PM), and sulfur dioxide. In addition, we will propose the use of air sensors connected with an app for citizens to manage the pollution level on a large scale. The remaining sections of this study are organized as follows. Section 2 is a literature review, which mainly states highlights and limitations from previous papers about air pollution in Valencia. Section 3 provides the main methods used in this project including gathering data, analyzing data, and surveying the citizens. Section 4 presents key findings of the work done, and section 5 discusses these results from the perspective of analysts. Finally, the study is concluded in section 6.

2 Literature review

2.1 Previous work

There have been several approaches to analyzing the trend of air pollution. For example, Betancourt-Odio et al. (2021) utilized the functional data and Kendall's Tau functional statistic (KFT) to identify significant correlations between areas in Madrid [3]. The team found that the ozone concentration is higher in rural areas due to the fact that places far from the center are more industrialized, or in other words, are used for building factories. Moreover, it was also seen that the level of O_3 starts to increase at 7 a.m. and reaches its peak at 3 in the afternoon, which later decreases to the lowest point at 9 p.m. In another paper conducted in Portugal a harbor city like Valencia, it is found that air pollutant dispersion relies on meteorological conditions such as atmospheric stability, wind direction, and sea-breeze circulation (Sorte et al., 2019) [4]. Additionally, during nighttime periods, the dispersion pattern is completely different from that in the daytime, and it promotes the accumulation of pollutants over the port area. Mateo Pla et al., on the other hand, focused on monitoring the road traffic data which is one the key elements regarding air pollution in modern cities [5]. The results show a highly detailed picture of GHG emissions in the city of Valencia with high temporal (hour) and space (street) resolutions.

Apart from monitoring and working with the electronic devices data in some of the papers authors measured public opinion acceptance towards proposed solutions that would reduce the level of air pollution in the cities. One of them was the paper of Saz-Salazar et al., (2020) in which public opinion towards alternative-fuel buses was presented [6]. The survey was carried out

and it turned out that 67% of the respondents are willing to pay extra for the adoption of this electric hybrid technology. This result shows that most of the citizens care about the air pollution problem in Valencia and are ready to take steps in helping with decreasing this issue.

2.2 Related Work

There are many papers that use various models to predict air pollution data and Cabaneros et al. collected all papers and drew graphs that show the neural network (NN) is the most popular and effective model among all [7]. Indeed, a multiplayer perception is used in more than 70 papers which yields great results in the prediction task, and the model training is mostly deterministic rather than stochastic. Besides, the common metrics for prediction tasks are variance, mean, and mix/max which present a whole picture of the air quality in a city. Considering other meteorological factors is important, so Contreras and Ferri (2016) use an interpolation method that takes wind direction into account, making the model more precise than the normal one [8]. Regarding the scope of the project, it is difficult to gather meteorological data in a short time, so we will instead focus on tuning the hyperparameters and metrics. Therefore, it is easier to collect and evaluate the data, and although the result will not be as good as those in the past, which is one of the main disadvantages of this paper, we expect to see the trend of air pollution in Valencia in the future and detect risk-potential areas. Future investigation into other factors influencing air quality is needed for better evaluation.

3 Framework

This paper presents a comprehensive approach to assessing air pollution in the city of Valencia using a combination of pollutant concentration data and citizens' opinion data. To collect the former, we utilize one main source of information: the Generalitat Valenciana website, which provides air quality data from sensors installed in various locations throughout the city. The data collected from these sources is numeric, representing pollutant concentrations in percentages, grams/milligrams/nanograms per cubic meter, among others. Using this data, we conduct a prediction analysis to identify areas of Valencia that are at risk for high pollution levels. By analyzing the concentration data, we will make informed predictions about the pollution levels in various regions of Valencia, which can be used to inform policy decisions by local authorities. Furthermore, we supplement this analysis with a survey of citizen opinions to gain a deeper understanding of their concerns about air quality and their thoughts on potential solutions. The survey data allowed us to analyze the citizens' perspectives on air pollution and assess their attitudes toward various potential interventions to address the problem. By integrating this data with our analysis of pollutant concentrations, we provide a comprehensive evaluation of the current situation of air pollution in Valencia and make targeted recommendations to local authorities to address the issue.

Regarding the proposed solution, at the moment, air sensors in cities are fixed because they are both very expensive and big to cover the entire city's air quality. However, this may overlook the fact that in most cases, these air sensors will mainly focus on their surroundings and ignore other places. Hence, we want to use small air sensors which are cost-effective to put in different parts of the city. To do that, we can offer citizens to buy accessories combined with the air sensor. As a result, when someone's surrounding is suddenly polluted because of unexpected reasons such as a big fire, ... the air sensor will detect these factors and send them to the authorities. Besides, air sensors in this case will not be fixed since people move to different places every day. Regarding the size of the air sensor, Bosch Sensortec has recently developed the world's smallest air sensor which provides timely, accurate, and actionable information about particulate concentration levels in the air [9].

The use of machine learning models to detect potentially polluted areas has emerged as an effective approach to addressing the issue of air pollution in urban areas. By leveraging machine learning algorithms, these models can identify patterns in air pollution data and make accurate predictions about areas that are likely to experience high levels of pollution. Once identified, local authorities can proactively address the problem, such as by reducing traffic congestion or implementing urban greening strategies like planting trees. This approach represents a promising long-term solution to the problem of air pollution, as it not only addresses the immediate problem but also focuses on the root causes of pollution. Additionally, conducting surveys on the issue of air pollution can provide valuable insights into the concerns and needs of citizens. By engaging with the public in this way, policymakers and local authorities can gain a deeper understanding of the specific issues that matter most to the community and design more effective interventions that meet their needs.

4 Methodology

4.1 Data Collection and Preprocessing

We have collated comprehensive data spanning the period from 2020 to 2023, gathered from 10 stations located in Valencia. However, there exist variations in the parameters measured across these stations. For instance, the Polytechnic station measures not only the concentration of PM10 but also PM2.5, while the Avd. Francia station measures wind velocity and direction. To resolve this issue, we have decided to prioritize the measurement of the most significant and prevalent pollutants in these stations, namely PM10 and NO2, because these ones are known to have significant impacts on human health.

In view of the concerns of the public, we conducted a small survey that sought to gather their opinions on air pollution in Valencia. The survey included questions that took into account the age of participants and their area of living, which we have included for the purpose of analysis. To encourage a higher level of participation, we ensured that the questions were short, straightforward, and concise.

4.2 Data Analysis Techniques

Firstly, to predict the pollutant concentration from the data, we first need to plot the pattern of the data and then decide which method we should apply. Based on the type of data, which is time series, a Recurrent Neural Network (RNN) is commonly used. We decide to use LSTMs (Long Short-Term Memory) which is a type of RNN architecture. LSTMs can handle time series data with variable-length input sequences, which is common in many applications. They can also effectively model long-term dependencies in the data, which is crucial for accurate predictions in many time series applications.

An LSTM unit typically consists of four components: a cell, an input gate, an output gate, and a forget gate. The cell is responsible for storing information for an indefinite amount of time, while the three gates manage the flow of information in and out of the cell. The forget gates determine which information from the previous state to discard by assigning a value between 0 and 1 to it, with 1 meaning the information is kept and 0 meaning it is discarded. Similarly, input gates determine which new information to store in the current state, and output gates control which information in the current state is output by assigning a value from 0 to 1 to it, based on both the previous and current states. By selectively outputting relevant information from the current state, the LSTM network can effectively maintain long-term dependencies and make accurate predictions for current and future time steps. We define some notation:

- x_t: input at time t
- h_t: hidden state at time t
- c_t: cell state at time t
- i_t, f_t, o_t: input gate, forget gate, and output gate activations, respectively, at time t
- W and U: weight matrices; b: a bias vector

Next, the formulas for LSTM are:

$$\begin{aligned}
\tilde{c}t &= \tanh(W_c x_t + U_c ht - 1 + b_c) \\
i_t &= \sigma(W_i x_t + U_i h_{t-1} + b_i) \\
f_t &= \sigma(W_f x_t + U_f h_{t-1} + b_f) \\
o_t &= \sigma(W_o x_t + U_o h_{t-1} + b_o) \\
c_t &= f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \\
h_t &= o_t \odot \tanh(c_t)
\end{aligned}$$

Here, $\odot$ represents element-wise multiplication, and σ represents the sigmoid activation function.

Regarding the survey data, we expect to gather around 50 responses from different ages and locations in Valencia to find some. This includes some statistics about their concerns and their evaluation of our proposed solution in

Valencia. They can also give constructive feedback so that we can adjust our solutions to be better.

4.3 Implementation Details

All the files of this project are found here: https://shorturl.at/bEIR9

5 Experiments and Results

5.1 Prediction results

The LSTM model demonstrates excellent performance and closely follows the trend of pollutant concentrations when trained on the training set and tested on the test set. This is supported by the low root mean square error between predicted and actual values. However, it should be noted that due to limited data availability in certain stations such as Olivereta, Ponen, Turia, and Vivers, the model's predictions cannot accurately capture the fluctuations in those locations. For instance, the Vivers station only has around 180 days of data for PM10 concentration measurement from 2020 to 2023, posing challenges for the LSTM model's prediction capabilities.

Given the satisfactory performance on the test set, we proceeded to utilize the model for predicting the next 20 days in Valencia. However, we encountered limitations in making longer-term predictions due to the presence of vanishing gradients. During training, as gradients are back-propagated through time, they may diminish, resulting in difficulties in capturing long-term dependencies. Moreover, as predictions extend into the future, errors and uncertainties in the predictions can accumulate and amplify, leading to decreased accuracy over extended sequences.

The obtained results are presented below. Although utilizing previously predicted data for making new predictions does not guarantee perfect accuracy, some notable changes in pollutant concentrations can still be observed. For instance, the NO2 concentration at Boulevard and Centre stations consistently exceeds 28 g/m^3, indicating potential future pollution issues in those areas. The Olivereta station exhibits highly fluctuating concentrations of both NO2 and PM10, ranging from 10 to 80 g/m^3. Despite the insufficient data for prediction, Turia, Ponen, and Vivers stations display typical fluctuation patterns in pollutant values over time. Additionally, the Sol station shows a gradual increase and decrease in NO2 concentration but a more pronounced variation in PM10 concentration. In contrast, the Francia station maintains a consistently high and stable PM10 concentration, while its pattern in the other pollutant appears to exhibit more intense swings. Lastly, despite significant differences in predicted trends for both pollutants, the Polytechnic and Pista de Silla stations share a similar long-term pattern in the future, with Pista de Silla having notably higher values.

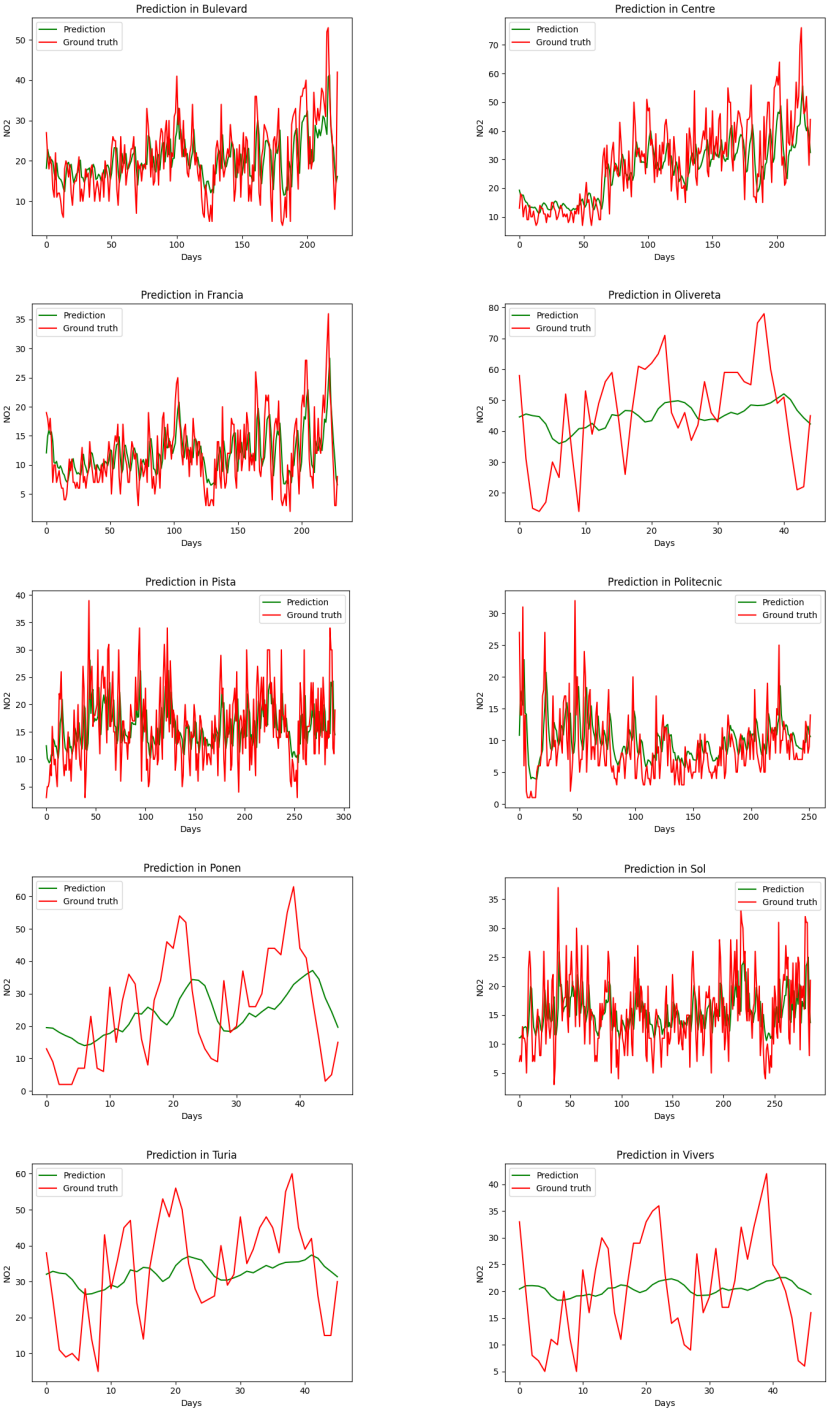

Fig. 1: Predictions on test data of NO2 concentration

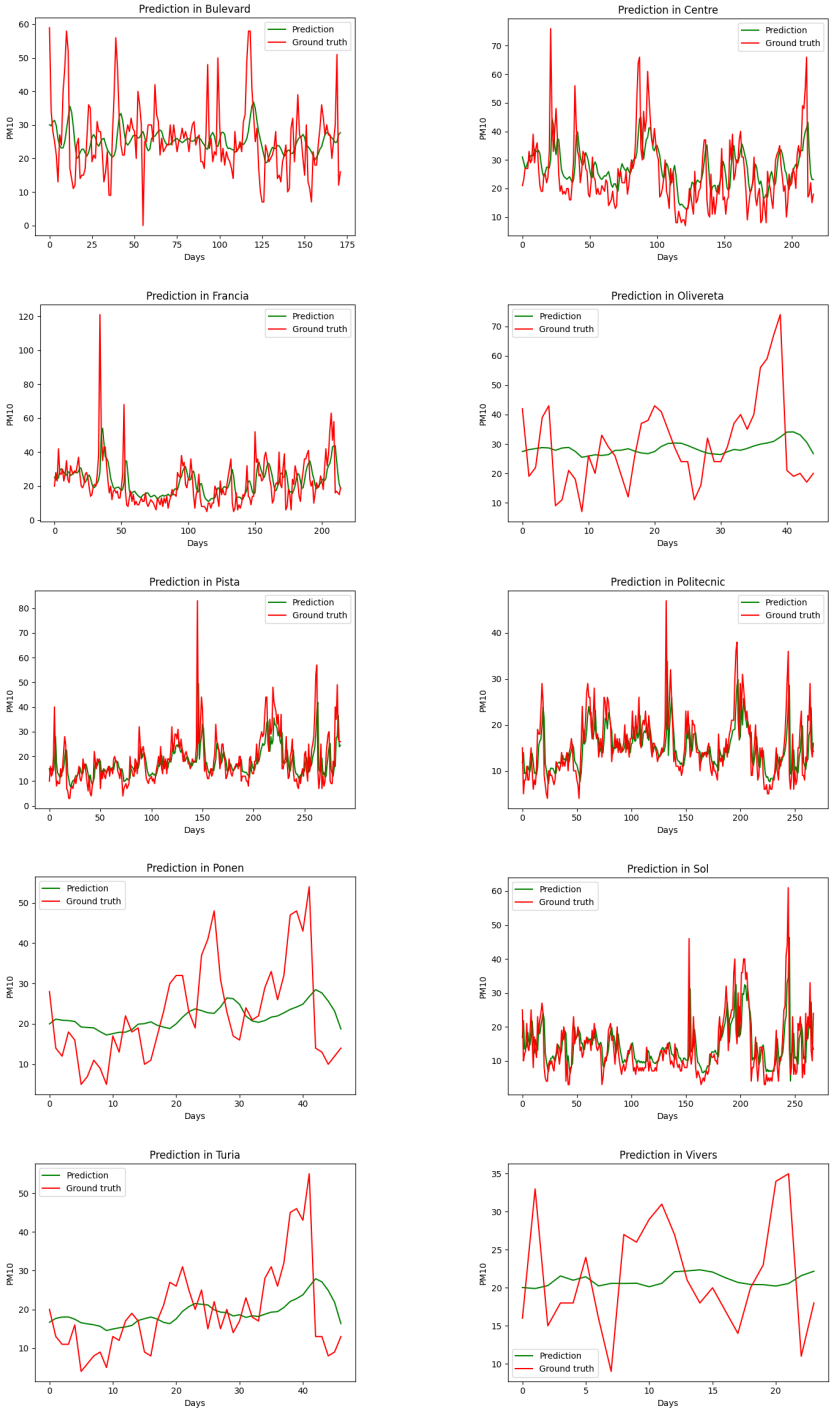

Fig. 2: Predictions on test data of PM10 concentration

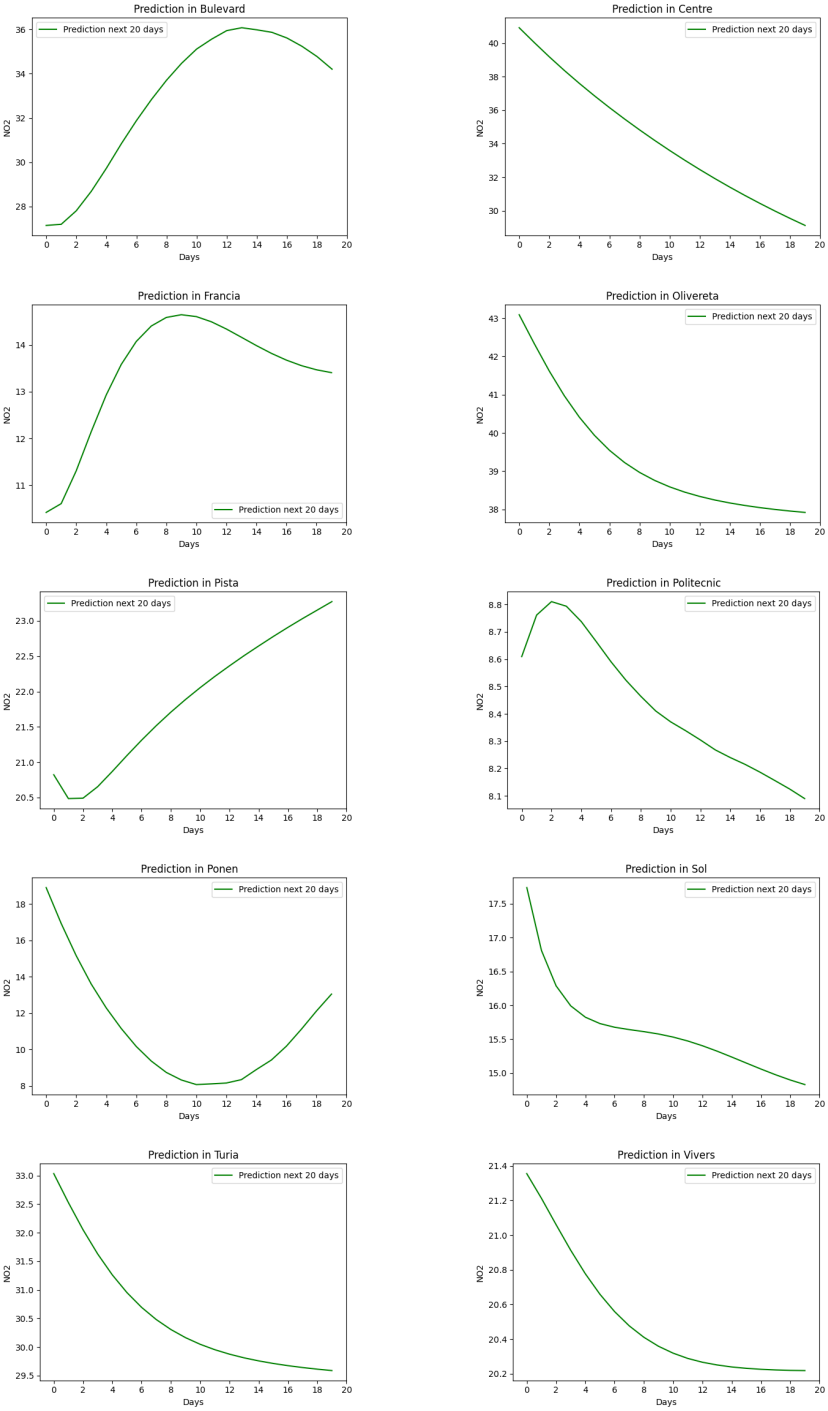

Fig. 3: Predictions of NO2 concentration in the next 20 days

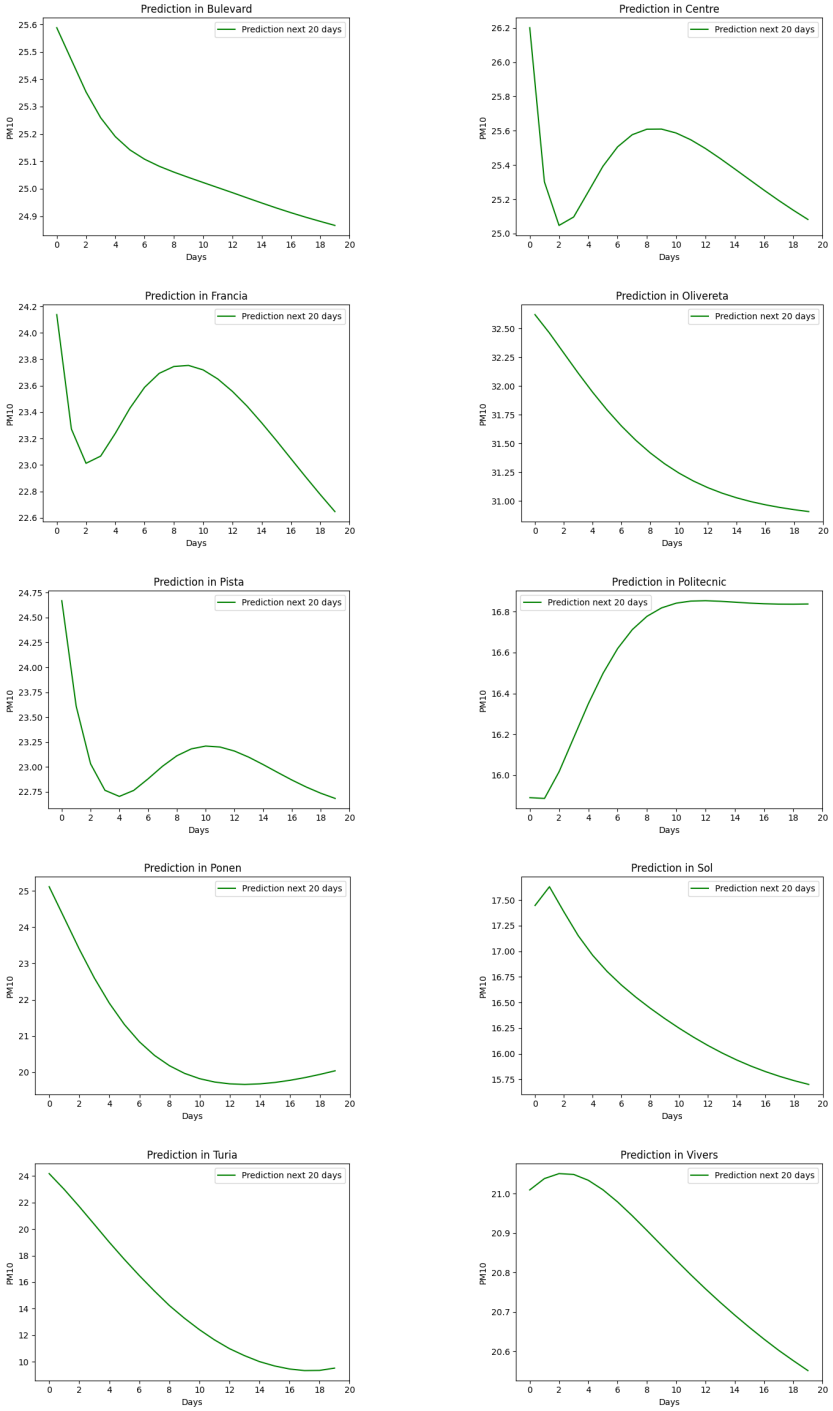

Fig. 4: Predictions of PM10 concentration in the next 20 days

5.2 Survey results

In regard to the survey, a total of 48 responses were obtained residing in the city of Valencia, providing their insights on a set of seven questions:

1. Which area in Valencia do you stay in?
2. You are in the age of?
3. On a scale of 1 to 10, how much do you rate the air quality in Valencia?
4. Do you have any concerns about the air quality in the area that you are living in?
5. Which one of these approaches do you think will significantly improve the air quality at the moment?
6. On a scale of 1 to 10, how much do you support our air sensor solution?
7. If you need to pay for the air sensor, at what price can you buy it?

The first five questions were focused on the general opinion of the citizens about air quality in Valencia. The last two questions were formulated to collect the opinion about our own air sensor idea. For questions number 3 and 6 the arithmetic mean has been used in order to calculate the average score of the answers provided by the participants in the survey. The formula for the arithmetic mean can be expressed as:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

With the appropriate notation for our case:

- x_i: combined score for each mark
- i: number of possible marks (in our case 10)
- n: number of participants

The number of responses is approximately equal to what we expected before, and although they cannot represent the whole population in Valencia, we still try to analyze them based on general measures such as locations and ages.

We received responses from participants residing in various districts within the city of Valencia, encompassing areas ranging from Cabañal to the outskirts. However, the highest number of responses was obtained from the districts of Blasco Ibanez, Cabañal, and Benimaclet.

With respect to the age of the participants, the majority of them are in their twenties, although there are also notable responses from individuals in the approximate age range of 40 years old, as depicted in the accompanying graph.

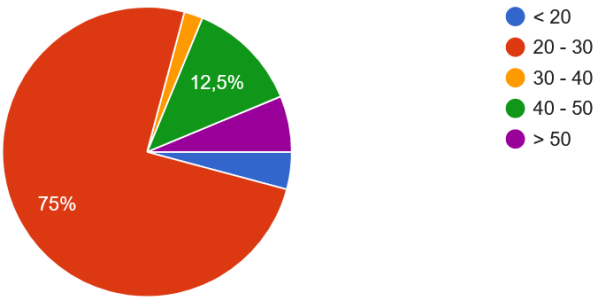

Fig. 5: The age distribution of participants

In the third question, the participants were asked to rate their experience on a scale ranging from 0, which corresponds to "Extremely bad," to 10, indicating "Excellent." The results of their answers are presented in below mentioned graph.

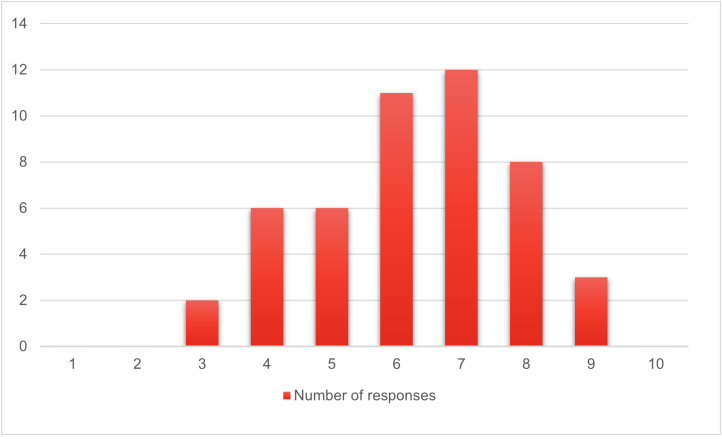

Fig. 6: Valencia air quality rating by participants

The analysis of the survey responses reveals a notable trend where the majority of participants favored scores 6 and 7. Moreover, upon calculating the arithmetic mean for these results, a score of 6.27 was obtained. This indicates that, according to the respondents, the air quality in Valencia is moderately satisfactory, implying that it does not rank among the worst. However, there remains room for improvement in order to enhance the overall air quality conditions.

When responding to inquiries regarding concerns about air quality in their residential areas, the student population in Valencia expressed varying viewpoints. While a subset of participants reported a few concerns, particularly during Fallas, another group residing in the same district did not observe any noteworthy air quality issues.

Then few improving air quality approaches have been presented for survey participants:

- Planting more trees
- Reducing traffic jams
- Reducing the smoke from factories
- Minimizing electricity use
- Utilizing a reduced quantity of fireworks

The responses can be seen below.

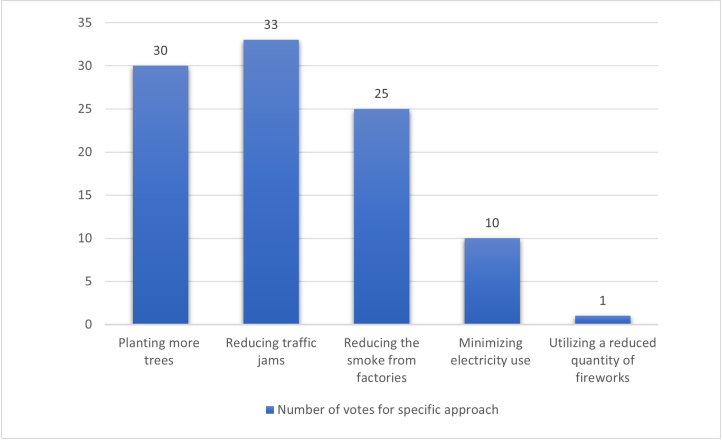

Fig. 7: Approach selection by participants

As observed from the aforementioned solutions, the implementation of measures with substantial support entails the expansion of tree-planting initiatives and the mitigation of traffic congestion. This signifies that the government of Valencia should prioritize the reconstruction of the urban traffic system in select areas, while simultaneously facilitating the establishment of additional green spaces and parks within the city, which presents a more feasible undertaking.

Following the initial set of general questions in the survey, we proceeded to introduce our air sensor solution to the participants, inquiring about their inclination to endorse the proposed solution using a scale ranging from 1 to 10. A rating of 1 denoted minimal support, while a rating of 10 indicated complete endorsement. The response was overwhelmingly positive, with 75% of the participants expressing a rating of 7/10 or higher. The average score obtained was

approximately 7.67. These findings underscore the substantial public support for our solution, suggesting a promising prospect for a wide customer base and future users. Consequently, further investment and continuous development in our solution are warranted. Below, you will find the graphical representation illustrating the exact scores of the responses.

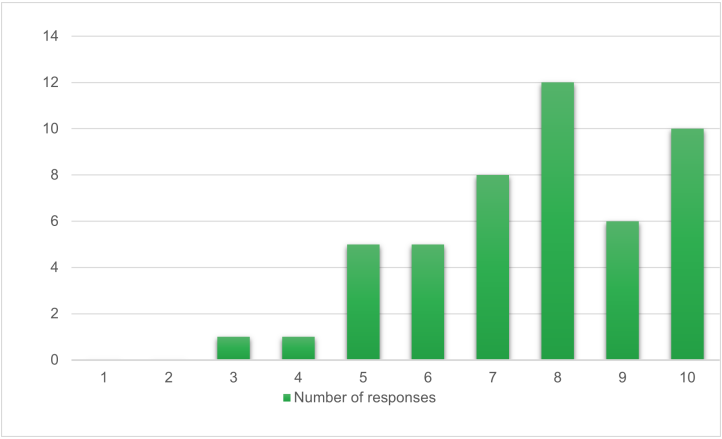

Fig. 8: Willingness of participants to support air sensor idea

Lastly, the matter of the air sensor's retail price was taken into account. Participants were queried regarding their monetary willingness to acquire the air sensor, and their responses are illustrated in the figure below.

Fig. 9: The participants' price preferences for the air sensor

The analysis of participant responses indicates a predominant inclination towards a maximum expenditure of 20 euros for the air sensor product. This pricing preference aligns reasonably with the product's size. While the exact production cost remains undetermined, it is unlikely to surpass 12 euros, suggesting a potential selling price of 15 euros, which is anticipated to satisfy

a substantial portion of the customer base. Moreover, it is essential to consider that a significant proportion of survey participants are in their twenties, indicating a stage of life characterized by limited financial resources and a propensity to allocate spending primarily toward essential items. Consequently, the inclusion of an older demographic in the survey might yield divergent outcomes.

6 Conclusion

In this paper, we have tackled two main goals including the use of the LSTM model to predict the concentration of two main pollutants, namely NO2 and PM10, and the analysis of Valencia's citizens about our proposed solution for air pollution in Valencia.

Firstly, regarding the prediction task, the LSTM model effectively predicts the concentration values of pollutants. The results show that the most potential areas that may be over-polluted in the near future are Boulevard and Center. However, one limitation of the model is the less accurate results when it predicts a point far from now due to the fact that the model uses its own prediction to generate new values, making big errors in long-term predictions. Therefore, future investigation into optimization techniques of the models such as forcing, scheduled sampling, or curriculum learning is needed.

The survey conducted in Valencia provided valuable insights into residents' perceptions of air quality and the proposed air sensor solution. Participants from diverse residential areas expressed moderate satisfaction with the current air quality situation, indicating room for improvement. Varied concerns about air quality were observed, with specific issues reported during certain periods. The proposed air sensor solution received substantial support, demonstrating its potential to address air quality concerns. Participants expressed willingness to invest in the solution at a reasonable price. Including participants from a wider age range would capture a broader range of opinions. The findings emphasize the importance of addressing air quality challenges in Valencia, including the expansion of green spaces and improvements in urban traffic management. Residents' strong commitment to improvement reflects their concern for well-being and living standards.

Acknowledgement

We would like to thank our professors Luján and Luis for their guidance, support, and valuable insights throughout the entire project. Their expertise and encouragement played a vital role in shaping the direction of this research and ensuring its quality. Indeed, we are provided with the necessary resources, facilities, and a conducive research environment, and our findings are instrumentally enriched by constructive feedback.

References

[1] Internations: The Cities Offering the Best (& Worst) Life Abroad (2022). https://www.internations.org/expat-insider/2022/best-worst-cities-for-expats-40327

[2] Briz-Redón, A., Belenguer-Sapiña, C., Serrano-Aroca, A.: Changes in air pollution during covid-19 lockdown in spain: A multi-city study (2020). https://doi.org/10.1016/j.jes.2020.07.029

[3] Betancourt-Odio, A., Valencia, D., Sofritti, M., Budría, S.: An analysis of ozone pollution by using functional data: rural and urban areas of the community of madrid (2021). https://doi.org/10.1016/j.envpol.2013.03.012

[4] Sorte, S.e.a.: Assessment of source contribution to air quality in an urban area close to a harbor: Case-study in porto, portugal (2019). https://doi.org/10.1016/j.scitotenv.2019.01.185

[5] Mateo pla, M.A.e.a.: From traffic data to ghg emissions: A novel bottom-up methodology and its application to valencia city (2021). https://doi.org/10.1016/j.scs.2020.102643

[6] Saz-Salazar, S., Feo-Valero, M., Vázquez-Paja, B.: Valuing public acceptance of alternative-fuel buses using a latent class tobit model: A case study in valencia (2020). https://doi.org/10.1016/j.jclepro.2020.121199

[7] Cabaneros, S.M., Calautitb, J.K., Hughes, B.R.: A review of artificial neural network models for ambient air pollution prediction (2019). https://doi.org/10.1016/j.envsoft.2019.06.014

[8] Contreras, L., Ferri, C.: Wind-sensitive interpolation of urban air pollution forecasts (2016). https://doi.org/10.1016/j.procs.2016.05.343

[9] BOSCH: World's Smallest Particulate Matter (PM2.5) Air Quality Sensor (450 Times Smaller in Volume than Alternatives) Enabling Use in Ultra-compact IoT Devices. https://www.bosch-sensortec.com/news/worlds-smallest-particulate-matter-sensor-bmv080.html