IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 12 October 2023, accepted 2 November 2023, date of publication 8 November 2023, date of current version 30 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3331565

RESEARCH ARTICLE

Data Driven Forecasting Models for Urban Air Pollution: MoreAir Case Study

SAFAA BERKANI[1], IHSANE GRYECH[1,2,3], MOUNIR GHOGHO[1,4], (Fellow, IEEE), BASSMA GUERMAH[1], AND ABDELLATIF KOBBANE[3], (Senior Member, IEEE)

[1]TICLab, International University of Rabat, Rabat 11103, Morocco
[2]Department of Electrical Engineering, KU Leuven, 3000 Leuven, Belgium
[3]ENSIAS, Mohammed V University in Rabat, Rabat 10102, Morocco
[4]Faculty of Engineering, University of Leeds, LS2 9JT Leeds, U.K.

Corresponding author: Safaa Berkani (safaa.berkani@uir.ac.ma)

This work was supported by the MoreAir Project, funded by the Belgium Ministry of Cooperation through the Vlaamse Interuniversitaire Raad-Universitaire Ontwikkelingssamenwerking (VLIR UOS) Program under Grant MA2017TEA446A101.

ABSTRACT Artificial Intelligence has the potential to contribute to sustainable cities, life on land, and climate action. Specifically, data-driven AI models can analyze large, interconnected databases to develop joint environmental actions. Air quality plays a pivotal role in both climate action and the development of sustainable cities, but developing countries face challenges due to insufficient monitoring stations and limited access to air quality data sets. This study builds upon the MoreAir project, which established a low-cost air pollution monitoring system and provided the first air quality data set from Morocco. We first exploit and delve into the details of the obtained dataset. Subsequently, we conduct a multi-level comparison of data-driven forecasting models, specifically focusing on short-term forecasting of Particulate Matter concentrations. Four forecasting frameworks are explored, using different combinations of exogenous data and spatio-temporal information. Our findings highlight that Machine Learning models, particularly LightGBM and CatBoost, outperform other models. Overall, our study demonstrates that the inclusion of the spatial dimension along with the diverse exogenous features enhances the models' predictive performance, and provides valuable insights.

INDEX TERMS Air pollution, urban air pollution forecasting, open datasets, statistical models, machine learning, deep learning.

I. INTRODUCTION

A. MOTIVATION

The escalating challenges of urbanization and economic development in recent decades have led to a rise in levels of air pollution in low- and middle-income countries, as reported by the World Health Organization [1]. The State of the Air 2020 Report highlights that Asia, Africa, and the Middle East endure the most alarming annual average exposure levels of fine particulate matter ($PM_{2.5}$) [2]. Fuller et al. reinforced the gravity of the situation, revealing that nearly 9 out of 10 people who die from pollution-related causes live in low- and middle-income countries [3].

The associate editor coordinating the review of this manuscript and approving it for publication was Michael Lyu.

Notably, Sub-Saharan Africa's low-income countries experience a considerable proportion of premature deaths and diseases attributed to pollution [4]. In Morocco, a comprehensive report titled "Toxic Air: The Price of Fossil Fuels," published by Greenpeace MENA, discloses that air pollution contributes to over 13,000 annual fatalities, accounting for nearly 7% of all deaths, placing it as the 8^{th} predominant cause of death [5].

These realities contrast starkly with global aspirations. The United Nation's 2030 Agenda has laid down "Sustainable Development Goals", with 17 goals and 169 targets aimed at achieving a range of improvements, from eradicating poverty and inequality, protecting the planet, to ensuring universal justice, prosperity, and health access. As part of these goals, three targets are addressing the air pollution

crisis, targeting reductions in related deaths and illnesses, ensuring access to clean energy in homes, and improving air quality in cities. However, the trajectory of air pollution remains concerning, necessitating worldwide substantive, sustainable interventions. To foster improvements in air quality, governments, researchers, and individuals are actively exploring diverse methods for monitoring, modeling, and forecasting air quality.

While air quality monitoring infrastructures have been established globally, Africa is severely underserved. A revealing statistic from UNICEF shows that a mere 6% of African children live within 50 kilometers of any air quality monitoring stations, in stark contrast to 72% in Europe and North America. Monitoring stations, despite their precise measurements, enable the capture of spatial variability, grapple with challenges. The prohibitively high cost and maintenance associated with deploying such networks in large numbers serve as significant barriers. Additionally, the placement of monitoring stations may not be optimal in proximity to areas of anthropogenic activities or high population densities, compromising the accuracy of the air pollution's spatial distribution estimations in urban regions, particularly in roadsides and major traffic congestion areas [6].

Recent endeavors have focused on the development of low-cost, compact sensors as means to address these challenges. Such sensors utilize cost-efficient components, data acquisition systems, and communication modules, and present an alternative to traditional air quality monitoring stations. Consequently, more open-air quality datasets have emerged, facilitating research in the field [7], [8], [9], [10]. Yet, the research landscape in Morocco remains scarce, primarily due to the lack of openly available air quality datasets.

This study bridges the knowledge gap by drawing insights from the first Moroccan air quality dataset collected using the MoreAir platform, as presented in previous work [12]. The MoreAir platform's design is adept at gauging outdoor air pollution originating from both diffuse and point sources of pollution. The selection of sites for this study ensures representation across diverse pollution origins and neighborhoods, setting the stage for accurate forecasting models.

The MoreAir dataset, named after the platform, provides context-specific information on air pollution, meteorological conditions, and exogenous measurements gathered from several neighborhoods in Rabat, the capital city of Morocco. Utilizing the distinctive MoreAir data, the objective is to deeply investigate the spatial, temporal, and contextual facets of Morocco's air quality. Another aim is to undertake an exhaustive assessment of several data-driven models across diverse forecasting frameworks and scenarios.

B. RESEARCH OBJECTIVES AND CONTRIBUTIONS

In response to the challenges and research gaps identified in prior studies, the current work presents a holistic, data-driven analysis for air quality prediction. It addresses a significant gap in air quality prediction, especially concerning African countries like Morocco, where there has been a historical dearth of openly accessible multivariate spatio-temporal air quality data. The main contributions of this work are summarized as follows:

- **Novel Dataset Introduction:** The MoreAir dataset is presented, making it the first of its kind for Morocco. Distinctively characterized by its diverse and multivariate input, this data set offers a pioneering opportunity for comprehensive air quality analysis in the region.
- **In-depth Data Analysis:** A deep exploration of the aforementioned data set is conducted, shedding light on air quality trends. The in-depth analysis is distinctive in that it uses a thorough methodology that considers temporal, spatial and context-specific elements, which have a substantial impact on influencing pollutant concentrations.
- **Comprehensive Model Evaluation:** This work stands out for its thorough evaluation of a range of forecasting models, spanning from traditional statistical methods to advanced deep learning techniques, such as mSSA, MLR, XGBoost, CATBoost, LightGBM, RF, SVR, LSTM, TCN, and MTGNN.
- **Diverse Forecasting Frameworks:** The study offers a thorough model evaluation and a unique comparison of predicting scenarios. Exogenous features and spatial data are incorporated into the forecasting models to assess their impact on the prediction accuracy. The findings underscore the marked improvement in prediction accuracy when both spatio-temporal and exogenous data are taken into account.

C. STRUCTURE

The structure of this paper is as follows. Section II provides a discussion on the study's background and initiates the data exploration, offering an extensive exploration of the MoreAir project and its corresponding dataset. This section describes the dataset attributes, the undertaken preprocessing steps, and the withdrawn insights specific to the Moroccan context. In section III, the methodology adopted for this study is unveiled, starting with the notations and problem formulation followed by an overview of the explored forecasting frameworks. Section IV elaborates on the evaluated models, the hyperparameter tuning process, and the performance assessment criteria used. In section V, a thorough presentation and analysis of the experimental results are provided. The paper concludes in Section VI, where it summarizes the findings, offers concluding remarks, and suggests potential directions for future research.

II. BACKGROUND AND DATA EXPLORATION

A. AI FOR AIR QUALITY

Artificial Intelligence (AI) has become an indispensable tool for addressing a plethora of modern challenges, including proactive monitoring and prediction of air quality. The gravity of air pollution, a global menace affecting health

and ecosystems alike, underscores the urgent need for innovative solutions. In this context, air quality prediction has become crucial in the pursuit of sustainable cities. The multivariate and intricate spatio-temporal patterns present in air quality data have necessitated the exploration of a diverse range of forecasting approaches, with data-driven forecasting techniques gaining significant traction [13].

From Statistical to Data-Driven Models: Historical attempts to predict air quality heavily relied on statistical approaches. For instance, the Multivariate Singular Spectrum Analysis (mSSA) [14], a non-parametric decomposition-based method, provided a solid statistical foundation, effectively tackling various forecasting scenarios due to its strong statistical underpinnings. However, with the increasing granularity and volume of data, and given the fact that it is challenging to intricate dynamics in air quality data, there was a clear need for more advanced techniques to capture complex non-linear relationships inherent in air quality data. This led to the gradual transition to machine learning models, and subsequently, deep learning models.

Machine Learning Models: Support Vector Regression **(SVR)** and ensemble tree-based models like Extreme Gradient Boosting **(XGBoost)**, Categorical Boost **(CatBoost)**, Light Gradient Boost Machine **(LightGBM)**, and Random Forest **(RF)** have pioneered this transition [15], [16], [17]. Their proficiency in modeling non-linearities made them the go-to choice for many researchers [18], [19], [20]. Concurrently, Multiple Linear Regression (MLR) [21] serves as a reliable tool for practitioners seeking a balance between model simplicity and interpretability.

Deep Learning Models: Recognizing the intricacies of air quality data, the research community shifted towards deep learning models, particularly those designed to capture long and short-term dependencies in the data [22]. Long Short Term Memory **(LSTM)** [23] networks, for example, exhibited significant prowess in understanding long-term patterns [24]. Conversely, Temporal Convolutional Networks **(TCN)**s [25], have been heralded for their efficiency in capturing short-term temporal relations, especially in Spatio-Temporal series where spatial interactions are pivotal.

Graph Neural Networks for Air Quality: The uniqueness of air quality data, characterized by interconnected influences from various sources, and the subsequent multi-variation of features, makes it apt for representation as a graph. Multiple works have exploited graph based models to address the forecasting task in air quality [26], [27], [28], [29]. The Multivariate Time Series Forecasting Graph Neural Network **(MTGNN)** [30] model stood out as it was one that yielded the most promising results. Its adaptability in learning the Spatio-Temporal graph coupled with its capability to perform convolutions across spatial and temporal dimensions makes it a powerful tool for such applications.

As technology continues to advance, leveraging its capabilities becomes vital for fostering a healthier and more sustainable environment. The complexity of air quality data combined with the scarcity of available datasets, especially multivariate spatio-temporal ones in regions like Africa, presents significant challenges. Addressing this, the research exploits the novel MoreAir dataset, generated by the MoreAir system. A range of models, from traditional to advanced ones, are engaged to offer a comprehensive perspective on air quality forecasting. This analysis highlights the unique strengths and weaknesses of each model, providing deep insights into the intricate challenge of air quality prediction. The overarching goal of our work is to furnish a thorough comparison across models and forecasting scenarios, ultimately contributing to the vision of sustainable cities with improved air quality.

B. THE MOREAIR PROJECT AND DATA SET

In our prior work [31], a novel approach was developed to create the MoreAir data set, aiming to examine the impact of multiple factors on air pollution levels. This unique data set was initially developed to investigate the potential relationship between environmental conditions and the respiratory health of asthma patients in their living environments. The main characteristics of the dataset are:

- **First in Africa:** The MoreAir data set is a novel achievement since its the first of its kind on the African continent. Its inception marked a momentous milestone in the domains of air quality monitoring and environmental health research within the region.
- **Comprehensive Data:** This dataset encapsulates three primary categories of environmental information: air quality data, weather data, and geographical data. This extensive collection spans time and encompasses diverse neighborhoods. It is composed of 52 temporal and geographical features, presenting a multidimensional perspective of the study areas.
- **Rich Geographical Features:** In addition to meteorological data, the MoreAir dataset comprises an array of geographical factors. These factors include an assortment of attributes across different zones within the study areas. The inclusion of these factors contributes to a nuanced comprehension of the environmental context.

To address concerns about the accuracy and reliability of the sensors used for air quality monitoring, an extensive evaluation was conducted in an earlier study [31]. Low-cost sensors, while cost-effective, often face challenges related to calibration, operational stability under various meteorological conditions, and intramodel variability. The evaluation confirmed that the sensors exhibited accuracy and repeatability, providing reliable measurements of particulate matter concentrations. Additionally, these sensors demonstrated the capability to detect events involving elevated particulate matter levels and identify PM "hot-spots." In a comparative experiment, the low-cost sensor was placed alongside the more sophisticated OPC-N3 optical particle counter from Alphasense, a sensor with well-established data quality validated in various studies. This experiment further reinforced confidence in the chosen PM sensor.

C. DATA SET DESCRIPTION

The unique MoreAir data set merges three different types of environmental data: air quality, meteorological, and geographical data.

Air quality and meteorological records: the data set consists of data reported by 3 air quality monitoring sensors settled in three locations with different characteristics in Rabat, Morocco. These air quality measurements were carried out using the MoreAir system [12]. The collected data is composed of measurements of fine particulate matters concentrations ($PM_{2.5}$ and PM_{10}), in ($\mu g/m^3$), temperature and relative humidity in °C and % respectively, as well as GPS data and timestamps associated with each measurement. Pollutant concentrations are recorded in real-time at five-second intervals over a three-month period, leading to over 966600 observations. This spans from November 11th, 2020, to January 6th, 2021.

Geographical Factors: the concentrations of air pollutants in urban areas arise from different sources, such as facilities, shops, and several activities. In our previous studies, these pollution sources were identified by monitoring the PM concentrations while walking through the narrow streets of the selected neighborhoods. Thus, two methods were used to extract spatial features: feature abstraction and micro-level scale data collection. With micro-level data collection, several context-specific pollution sources were identified, addressing the limitations faced when using feature abstraction. This led to the acquisition of 52 factors, mainly: Neighborhood, Buffer Size, Zone, Distance to nearest road, Distance to nearest large road, Type of nearest road, Distance to nearest Public Bath/ Oven, Distance to nearest Green area, Distance to water area, Water area Type, Green area Type, Distance to nearest Bus Station, Distance to nearest Taxi Station, Distance to nearest train station, Distance to tramway station, Distance to nearest Craft, Distance to nearest traditional Market, Distance to street vendors, Distance to Cuisine, Street Food, Cuisine, Commercial, Crafts, Street Vendors, Commercial market, Commercial Center, School, Public Establishment, Public baths, Public Ovens, Tourism, Hospital, Tramway stops, Bus Stops, Taxi Stops, Sport, Industry, Parking, Green Land area, Water Area, Arterial Roads, Low capacity roads, Dual Carriageway, Roundabout, Tunnels, Parkways and Height of sampling point.

We made use of geographical factors during the data exploration phase. These factors were instrumental in understanding spatial variations and their potential impact on the analysis. While they were primarily used for data exploration, they played a crucial role in gaining insights into the dataset's spatial characteristics and ensuring the robustness of the employed models' predictions. In the experimental section, we focused on the main exogenous data, such as meteorological factors like humidity and temperature, which directly influenced the model's predictions. This comprehensive approach allowed us to build a solid foundation for the research and produce reliable results.

One of the primary objectives of the MoreAir project is continuous air quality data collection. Thus, the sensor nodes have been actively upgrading to enable uninterrupted measurements and gather new data [32]. This ongoing data collection effort is crucial for advancing our field research, enhancing predictive capabilities, and providing a valuable resource for future studies in this domain.

D. DATA SET PRE-PROCESSING

Handling missing or noisy data from the sensors was a critical aspect of this study's data pre-processing. Given the dependence of Spatio-Temporal forecasting on both spatial and temporal granularity, a meticulous approach was followed to ensure data quality and reliability. The key steps taken to address these issues include:

- **Spatial Coverage:** Due to the installation of sensor nodes at different locations, alignment issues arose during the data collection process. Therefore, our data pre-processing began with a spatial analysis to determine the spatial coverage accurately. Three distinct neighborhoods, with unique characteristics, were chosen to represent the geographical diversity in the data. The first neighborhood is considered one of the city's most opulent areas, distinguished by expansive roads, abundant green spaces, and modern architecture. The second neighborhood is one of the oldest in the city, with negligible traffic but a dense population and diverse traditional activities. Additionally, it is situated in close proximity to the beach, an important natural source of particulate matter pollution. The third neighborhood shares similarities with the second in terms of population and daily activities but exhibits distinct spatial characteristics.
- **Temporal Coverage:** To further enhance data quality, the temporal span was limited to the period when all three sensors were actively collecting and transmitting data. The selected air quality time series was then aggregated into 15-minute intervals using the arithmetic average, guaranteeing consistent and continuous temporal coverage.
- **Denoising the Dataset:** The careful selection of both spatial and temporal spans, coupled with the arithmetic average aggregation along the temporal axis, significantly reduced the presence of missing data. However, as part of the meticulous preprocessing, the issue of noisy input data was also addressed through an outlier detection process. The 95% quartile outlier detection technique was implemented to identify and filter out noisy data points. These outliers, often caused by sensor inaccuracies or anomalies, were carefully identified and addressed to enhance data quality.
- **Missing Data Imputation:** Since the number of missing values remained relatively low after these preprocessing steps, missing data imputation was carried out using a moving average model, implemented with the imputeTS library in (**R**) [33]. This step ensured that any remaining

gaps in the data were effectively filled, maintaining data integrity.

The combined effect of these data handling procedures was to ensure that the dataset was of high quality, free from excessive noise, and to effectively address any missing data. These steps were taken to enhance the reliability of the analysis and the accuracy of the findings.

Following data pre-processing, the dataset was transformed into a 3D-Tensor. The first axis, with **5371** observations, accounts for time. The second axis encapsulates the spatial dimension, covering **3** distinct spatial points. Lastly, the third axis represents both endogenous and exogenous factors, consisting of **4** variables.

The dataset was divided into two subsets: 80% for training and 20% for testing. The time series splitting method was employed to prevent information leakage. Prior to feeding the data into the forecasting models, it was normalized using a feature scaling technique. To ensure a fair comparison, all models are subjected to nearly identical assessment approaches. Thus, the training and testing sets remain fixed across all the applied models.

E. DATA SET EXPLORATION

In this section, the temporal variations in the distribution of air quality is investigated and the impact of spatial features is examined.

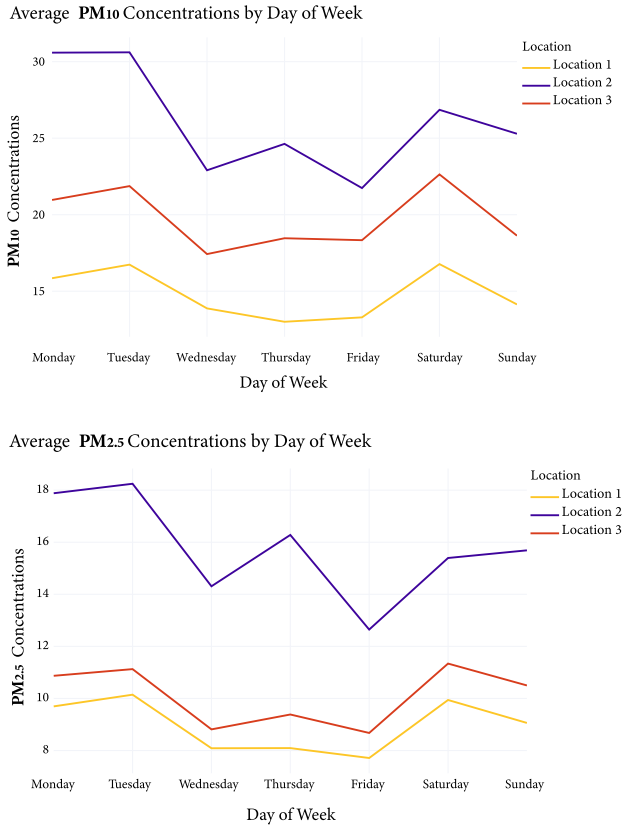


FIGURE 1. Variation of $PM_{2.5}$ and PM_{10} records by day of week.

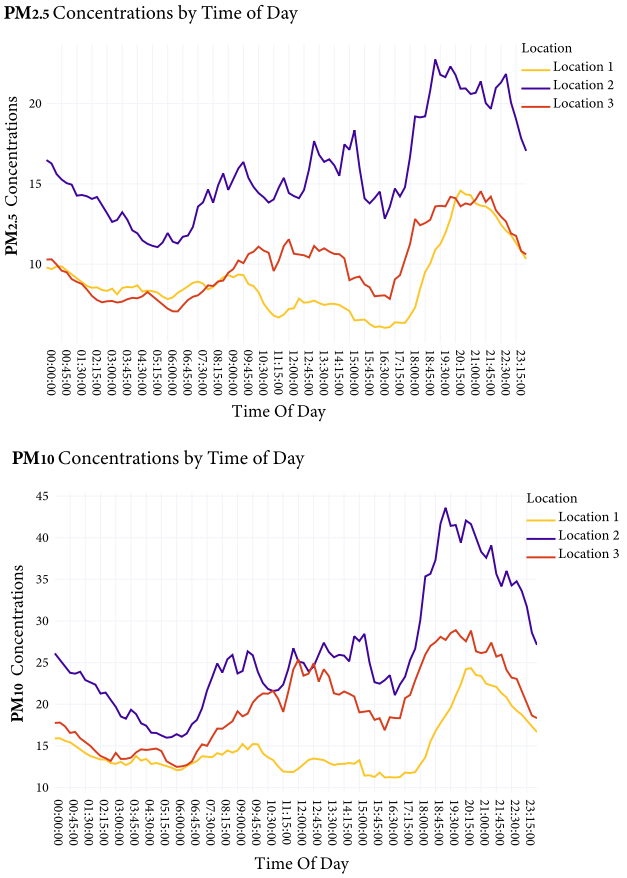


FIGURE 2. Variation of $PM_{2.5}$ and PM_{10} records by time of day.

Fig.1 presents the average concentrations of $PM_{2.5}$ and PM_{10} for each day of the week. It is observed that both pollutants exhibit significantly higher concentrations on Monday, Tuesday, and Saturday across most locations. The lowest values were recorded on Wednesday and Friday.

Fig.2, depicts the average concentrations of $PM_{2.5}$ and PM_{10} throughout the day in the three studied locations. Although the concentrations vary on different scales, both pollutants display similar fluctuations. Clearly, their mean concentrations gradually increase and reach their peaks towards the end of the day. Additional peaks are observed during the morning hours (08:00 to 10:00) and the afternoon hours (13:00 to 16:00), which align with rush hours when people commute to and from work or school.

Even when considering daily concentrations in detail, Location 2 consistently exhibits the highest pollutant concentrations, whereas the two other neighborhoods share low values. The similarity between Location 1 and Location 3 is particularly evident from midnight to 08:00, before diurnal activities commence.

Comparing the dynamics of pollutant concentrations across different locations enables empirical investigations into their spatial distribution in areas with distinct characteristics.

Fig.3 shows the number of pollution sources or impacting factors identified in the vicinity of the studied locations. Location 2 has the highest number of impacting factors, followed by location 3, which can be attributed to its high population and diverse range of traditional activities.

In Fig.4, the y-axis displays the closeness of the studied locations to the different pollution sources and impacting features. It is evident that locations 2 and 3 are significantly closer to traditional markets, crafts, and water areas. Both the proximity to pollution sources and the number of sources contribute to the similar variation in PM concentrations observed in Locations 2 and 3. However, given its proximity to the beach, Location 2 experiences significantly higher records, as it is exposed to sea salt, which is a natural source of particulate matter. Conversely, Location 1, representing the opulent neighborhood, with the fewest pollutants and farthest from crafts and traditional markets, records the lowest values.

The conformity between the results observed in Figures 1 and 2, and those observed in Figures 3 and 4, confirms the entanglement between the temporal and spatial dimensions, which motivates our exploration of the third and fourth frameworks of our application.

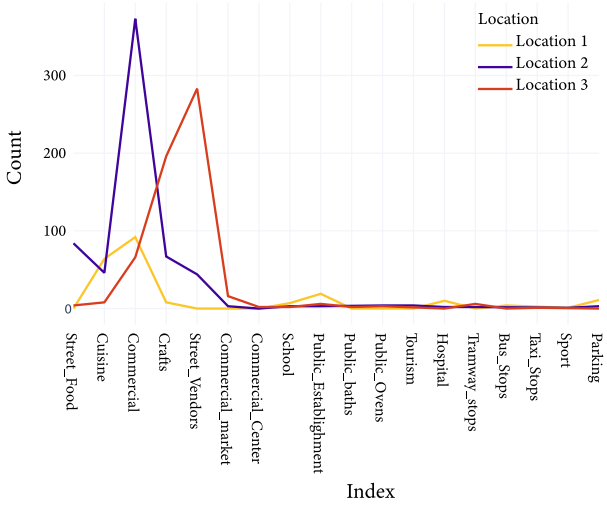


FIGURE 3. Number of sources contributing to air pollution by location.

III. METHODOLOGY

A. NOTATIONS AND PROBLEM FORMULATION

In this paper, the task of collective, single-step-ahead forecasting is addressed. Where the objective is to collectively predict the future values of one target variable, denoted as Y, among m variables collected at s locations at n discrete points in time. This is represented by the data tensor $\mathcal{Z} \in \mathbb{R}^{m\times s\times n}$. Given the historical window of length p from time step $t-p+1$ to t, represented as $\mathcal{Z}(t-p+1:t)$, where

$$\mathcal{Z}(t-p+1:t) = (\mathbf{Z}(t-p+1), \ldots, \mathbf{Z}(t-1), \mathbf{Z}(t)) \quad (1)$$

The aim is to build a mapping $f(.)$ that takes the historical data as input and yields the forecast of the target variable at horizon 1, denoted as $y(t+1) = \left(y^{(1)}(t+1), \ldots, y^{(s)}(t+1)\right)$.

Here, $\mathbf{Z}(i) = \left(\mathbf{Z}^{(1)}(i), \ldots, \mathbf{Z}^{(s)}(i)\right)$ is a matrix given by:

$$\mathbf{Z}(i) = \begin{pmatrix} y^{(1)}(i) & \ldots & y^{(s)}(i) \\ z_1^{(1)}(i) & \ddots & \vdots \\ \vdots & \ddots & z_{m-2}^{(s)}(i) \\ z_{m-1}^{(1)}(i) & \ldots & z_{m-1}^{(s)}(i) \end{pmatrix} = \begin{pmatrix} y(i) \\ z_1(i) \\ \vdots \\ z_{m-1}(i) \end{pmatrix} \in \mathbb{R}^{m\times s}$$
$$\forall i \in \{1, \ldots, n\}.$$

B. FORECASTING FRAMEWORKS OVERVIEW

In this paper, the forecasting problem is examined under four distinct frameworks. These frameworks are designed to capture distinct aspects of the forecasting task while evaluating the impact of various features on the forecast accuracy:

1) TIMES SERIES (TS) FRAMEWORK

Within this framework, historical pollutant concentration data for both $PM_{2.5}$ and PM_{10} from the three monitoring locations are independently utilized as input features. The primary objective is to forecast pollutant concentrations at each location based solely on their respective historical values. This framework serves as a foundational baseline for comparisons with more intricate frameworks, facilitating a clear evaluation of the employed forecasting models. During the experiments, each location is considered individually, and only the target time series is incorporated into the forecasting model. Each neighborhood is treated separately then the overall performance is reported. The predicted values are computed using the mapping function as described in Equation (2).

$$\hat{y}^{(j)}(t+1) = f\left(\mathbf{y}^{(j)}(t-p+1:t)\right), \quad \forall j \in \{1, \ldots, s\} \quad (2)$$

2) TIME SERIES WITH EXOGENOUS DATA (TSX) FRAMEWORK

Building upon the TS framework, additional exogenous features are introduced, such as temperature and humidity, alongside both $PM_{2.5}$and PM_{10} as input variables. The aim here is to assess how contextual variables impact forecasting accuracy. The output values for each location, are obtained as described in Equation (3).

$$\hat{y}^{(j)}(t+1) = f\left(\mathbf{Z}^{(j)}(t-p+1:t)\right), \quad \forall j \in \{1, \ldots, s\} \quad (3)$$

It's important to note that in both TS and TSX frameworks, each location is treated independently without considering relationships with other locations.

3) SPATIO-TEMPORAL (ST) FRAMEWORK

In this framework, the spatial dimension is explored by utilizing data from all three monitoring locations simultaneously. This framework seeks to capture spatial correlations and dependencies between locations when forecasting pollutant concentrations. The input incorporates historical pollutant data from all monitoring locations. Throughout the experiments, two distinct and independent scenarios are explored: one centered on predicting $PM_{2.5}$ concentrations and another on

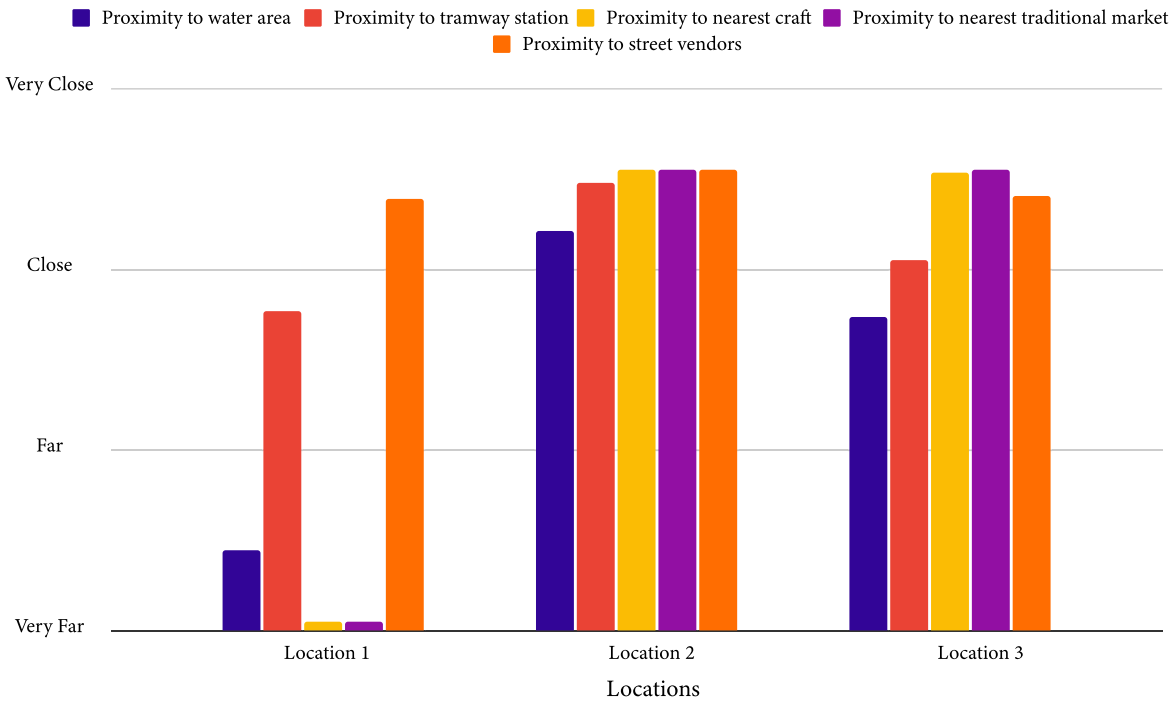


FIGURE 4. Distances to impacting factors by location.

forecasting PM_{10} concentrations. The final collective output is determined using Equation (4).

$$\hat{y}(t+1) = f\left(\mathbf{Y}(t-p+1:t)\right) \tag{4}$$

4) SPATIO-TEMPORAL WITH EXOGENOUS (STX) FRAMEWORK

Extending the ST framework, the STX framework includes exogenous features, mirroring the TSX framework. Here, spatial information from all monitoring locations is combined with contextual variables. The objective is to evaluate how effectively models can leverage both spatial and contextual data for forecasting. The final forecast is computed using Equation (5).

$$\hat{y}(t+1) = f\left(\mathcal{Z}(t-p+1:t)\right) \tag{5}$$

By considering the four distinct frameworks, summarized in table 1, the impact of incorporating the spatial dimension and exogenous variables on the accuracy of the forecasts can be analyzed and evaluated. The comparative analysis provides insights into the effectiveness of contrasted models under various scenarios, highlighting the importance of spatial dependencies and exogenous information in improving forecasting performance. In the experiments, it should be noted that when forecasting $PM_{2.5}$ and PM_{10} independently in the TS and ST frameworks, they are treated as unrelated tasks. However, in the TSX and STX frameworks, when forecasting one pollutant, the other serves as additional exogenous data to enhance the forecasting accuracy.

TABLE 1. Summary of the four forecasting frameworks.

	Temporal	Spatial	Exogenous Data
TS	✓	✗	✗
TSX	✓	✗	✓
ST	✓	✓	✗
STX	✓	✓	✓

IV. EXPERIMENTAL SETTINGS

A. ASSESSED MODELS

In this section, the ten algorithms selected for the comprehensive comparative study are introduced. The careful choice of data-driven forecasting models was driven by several pivotal considerations:

Diversity and Coverage: The objective is to encompass a wide spectrum of data-driven forecasting techniques, ensuring a comprehensive assessment of air quality prediction. The selection includes statistical, traditional machine learning, and deep learning models, providing a holistic view of their performance in the context of air quality forecasting.

Real-World Applicability: The chosen models represent a blend of well-established and state-of-the-art techniques commonly applied across various forecasting domains. This reflects their practical relevance and makes the findings valuable to both researchers and practitioners involved in air quality prediction.

Baseline Comparison: To establish a benchmark, simple models like MLR and mSSA are included. These models serve as reference points for evaluating the performance of more complex methods and assessing whether the added complexity of advanced techniques yields significant improvements.

Machine Learning Advancements: Machine learning models are incorporated, including XGBoost, CatBoost, LightGBM, and Random Forest, due to their demonstrated capability to capture intricate non-linear relationships within data. These models have achieved groundbreaking results in various data-driven tasks, rendering them suitable for handling the intricate patterns in air quality data.

Deep Learning Relevance: To assess their suitability and performance alongside traditional approaches, deep learning models like LSTM, TCN, and the MTGNN are employed. DL models are especially pertinent in scenarios where temporal and spatial relationships play crucial roles. LSTM and TCN were initially applied in time series frameworks and subsequently extended to incorporate spatial dimensions when applicable. MTGNN, due to its multivariate nature, found utility in the third and fourth frameworks.

By meticulously considering these factors, the goal was to conduct a comprehensive evaluation, catering to various aspects of air quality forecasting and ensuring the practical relevance of the study.

B. HYPERPARAMETER TUNING

The precision of forecasting models is intrinsically linked to the choice of hyper-parameters. To ensure the comparison between models is fair and impartial, a rigorous grid search strategy was utilized for each model. This aimed to uncover the optimal combination of parameters that would enhance their predictive efficacy when applicable.

For instance, to tune the mSSA model, ranks were adjusted between 5 to 20, segment lengths were experimented with, and choices were alternated between normalization and direct variance options. For the machine learning models, an array of learning rates from 0.001 to 0.1 was tested and diverse loss functions such as RMSE, MAE, and RMSLE were incorporated when relevant. For XGBoost, configurations spanned different booster types, including DART (Dropouts meet Multiple Additive Regression Trees), gblinear, and gbtree. Diverse learning objectives, such as regression with Pseudo Huber loss, squared loss, and squared log loss were also explored. LightGBM's fine-tuning consisted of adjustments in boosting types: traditional Gradient Boosting Decision Tree, DART, and Gradient-based One-Side Sampling. Meanwhile, the SVR adjustments focused on kernel types—radial basis function, linear, and polynomial kernel—with epsilon values tweaking from 0 to 0.01. Regarding TCN, LSTM, and MTGNN, the number of layers, filters, and learning rates were adjusted. The impact of incorporating dropout was also evaluated. For LSTM, unit numbers were an added variable, while for TCN, dilation rates were included in the tuning process. It's important to stress that these configurations are critical to ensure that the models successfully strike a balance between recognizing complex data patterns and avoiding overfitting, even though they give a peek of the substantial tuning effort employed.

C. PERFORMANCE ASSESSMENT

The evaluation of model performance in our study relies on two key metrics: the coefficient of determination (R^2) and the Mean Squared Error (MSE), defined as follows:

$$R^2 = 1 - \frac{\sum_{j=1}^{n} \left(Y_j - \hat{Y}_j\right)^2}{\sum_{j=1}^{n} \left(Y_j - \bar{Y}\right)^2} \tag{6}$$

$$MSE = \frac{1}{n} \sum_{j=1}^{n} \left(\hat{Y}_j - Y_j\right)^2 \tag{7}$$

where n represents the size of the test set, Y_j denotes the j-th observed value, and $\hat{Y}_j$ signifies the corresponding predicted value.

To ensure a robust assessment, five separate runs of the models were conducted, and their average errors compared. Given that this study examines different locations independently in the TS and TSX frameworks, the overall performance of each model is reported.

The chosen data-driven models are used to forecast one-step-ahead concentrations for both $PM_{2.5}$ and PM_{10}. Experiments utilize historical data gathered from three locations (s=3) with a window size of 14-time stamps (p=14).

V. RESULTS

This section evaluates the performance of 10 various forecasting models tested on the two pollutants of interest: $PM_{2.5}$ and PM_{10}. Table 2 and table 3 represent the different frameworks, and summarize the models' performance depicted by the values of R^2 and MSE.

It is shown that within the time series' framework, XGBoost outperforms the contrasted models for both pollutants. Compared to the persistent model, XGBoost achieves a decrease in MSE of 15.98% for PM_{10} and of 8.97% for $PM_{2.5}$. XGBoost scores the highest R^2 (0.71) for $PM_{2.5}$, followed by LSTM (0.70) and LightGBM (0.69). Similarly, for PM_{10}, XGBoost outperforms Catboost with an R^2 of 0.66, and LighGMB with an R^2 of 0.65, as well as the other employed forecasters.

The low performance of the deep learning methods, LSTM and TCN, as reflected by their high MSE and low R^2 scores, can be attributed to the non-linearity of the pollutants data along with the limited number of data samples. In contrast, the boosting models demonstrate their ability to capture dynamics even with relatively small data sets.

In the second framework, when exogenous variables are introduced, CatBoost and LightGBM show improved performance for both pollutants, making them the leading forecasters. LightGBM, for instance, achieves a decrease in MSE of 3.21% and 6.62% for PM_{10} and $PM_{2.5}$ respectively across the frameworks. However, XGBoost and Random Forest failed to improve due to the increased number of input

TABLE 2. *Time series forecasting comparison. the best results are shown in bold.*

Model	PM_{10}				$PM_{2.5}$			
	TS		TSX		TS		TSX	
	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2
XGBoost	**14.0387**	**0.6905**	16.1539	0.6072	7.0754	**0.7112**	7.2014	0.6823
CatBoost	14.9005	0.6678	14.9286	0.6566	7.2266	0.6976	**6.7501**	**0.7208**
Light GBM	15.0654	0.6599	**14.5818**	**0.6708**	7.2740	0.6998	6.7921	0.7144
SVR	16.5269	0.6441	15.9113	0.6545	7.8251	0.6781	7.7508	0.6818
Linear	16.9616	0.6194	16.3416	0.6333	7.3702	0.6960	7.3475	0.7001
Random Forest	16.2757	0.6039	17.7403	0.5240	7.9258	0.6632	7.3716	0.6694
LSTM	16.5014	0.5903	15.9969	0.6347	**6.9960**	0.7047	7.227	0.6694
TCN	15.8625	0.6398	16.3094	0.6278	7.0650	0.6783	7.5092	0.6842

TABLE 3. *Spatio-temporal forecasting comparison. the best results are shown in bold.*

Model	PM_{10}				$PM_{2.5}$			
	ST		STX		ST		STX	
	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2
XGBoost	15.1683	0.6652	19.4664	0.5590	7.1002	0.7046	8.832	0.6100
CatBoost	14.7676	0.6861	13.6584	0.6969	6.8435	0.7221	**6.2992**	**0.7396**
Light GBM	14.8907	0.6794	**13.3716**	**0.6984**	6.7063	**0.7344**	6.6389	0.7365
SVR	16.1827	0.6632	15.3161	0.6648	7.4013	0.6902	7.1995	0.7003
Linear	17.0007	0.6104	18.0479	0.5574	7.5657	0.6756	7.9559	0.6572
Random Forest	18.735	0.592	19.3075	0.5793	9.1503	0.6258	8.5178	0.6448
mSSA	18.3203	0.6079	18.2117	0.6063	8.6041	0.6699	8.3534	0.6772
LSTM	15.8303	0.6348	15.4092	0.6376	7.1850	0.6812	7.2226	0.6951
TCN	17.5314	0.5695	31.6034	0.3724	7.5640	0.6217	9.2995	0.4437
MTGNN	**13.6397**	**0.6975**	14.2031	0.6936	**6.6247**	0.7342	7.0827	0.7378

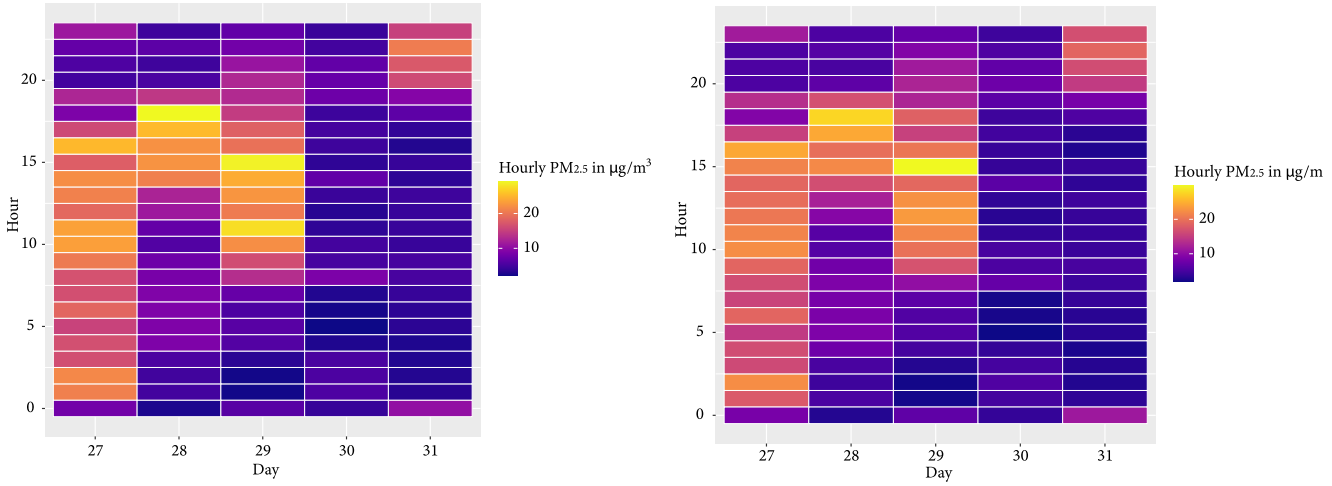
features compared to our limited number of entries in the data, which led to an extrapolation flaw in these tree-based models.

In the third framework, where Spatio-Temporal data were provided, table 3 shows that the graph-based MTGNN model achieved a higher R^2 score compared to the contrasted models for both pollutants. The satisfactory results of MTGNN, despite the relatively limited entries in our data set, reflect the model's capacity to depict spatial and temporal fluctuations simultaneously. In terms of forecasting performance, CatBoost and LighGBM ranked first.

These boosting models show further improvement when exogenous data are incorporated, surpassing the performance of the MTGNN for both pollutants. Specifically, LightGBM and Catboost achieve an R^2 of 0.698 and 0.696, and an MSE of 13.371 and 13.658, respectively. Therefore, these two models perform the best across all frameworks for predicting both $PM_{2.5}$ and PM_{10} concentrations.

The diversity of the forecasting scenarios allows us to deduce that incorporating both exogenous data and the spatial dimension led to significant improvements in the performance of the employed models. For instance, Catboost applied on the Spatio-Temporal data with exogenous features reduces the persistent model's MSE by 18.25% for PM_{10} and by 18.96% for $PM_{2.5}$. The inclusion of exogenous features in our data set provides clear insights into the dynamics of the data and enhances the performance of the forecasters by adding more structure to the input.

However, XGBoost and Random Forest are negatively affected by the incorporation of the spatial dimension and exogenous data, mainly due to the extrapolation flaw inherent in these tree-based models. It is worth mentioning that our study reveals that LightGBM and CatBoost, as machine learning models, are most suitable for addressing the formulated problem across the different frameworks.



(a) Real $PM_{2.5}$ measurements over time.

(b) Forecast $PM_{2.5}$ measurements over time.

FIGURE 5. **Comparison between the forecast and real values for $PM_{2.5}$.**

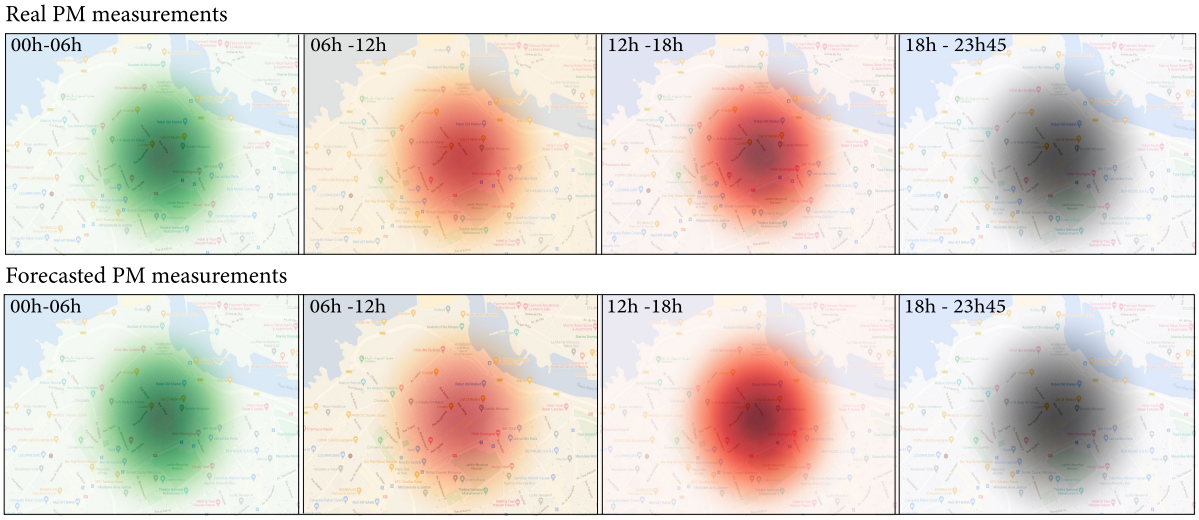


FIGURE 6. **Geographical heatmap.**

The forecast $PM_{2.5}$ concentrations averaged hourly, depicted in Fig.5b, visually validate the predictive performance of the MTGNN forecaster. Specifically, when contrasted to the real values displayed in Fig.5a, we clearly see that the forecaster captured the peak hours, all while seizing the main variations within the data.

In Fig.6, we present the geographical heat maps of PM concentrations. When contrasting the heatmaps generated by MTGNN with the real values' heatmaps, we notice that MTGNN clearly captures the average concentration levels throughout the day. In conformity to the average concentration of pollutants depicted in fig.2, the highest concentrations are observed at night (18h - 23h45).

VI. CONCLUSION

This study is a sequel to the MoreAir study, which introduced a low-cost urban air pollution monitoring system, creating a novel and unique context-specific air quality data set from Morocco. Our initial exploration of this dataset involved a multi-level comparison of various data-driven models to address the problem of short-term forecasting of Particulate Matter concentrations. The experiments were conducted under four different forecasting scenarios, using statistical methods, machine learning models, and deep learning approaches. Among the assessed models, ML algorithms, particularly LightGBM and CatBoost, emerged as leading performers in three of these frameworks. Specifically, LightGBM achieved the highest R^2 and lowest MSE in the PM_{10} TSX and STX frameworks. In contrast, CatBoost dominated the same frameworks for $PM_{2.5}$ predictions, presenting optimal R^2 and MSE metrics. The deep learning MTGNN model, despite the constraints of a limited dataset, demonstrated satisfactory results in the ST framework for both $PM_{2.5}$ and PM_{10}. Additionally, the findings underscore the enhanced forecasting

precision achieved by incorporating the spatial dimension and exogenous features into the dataset.

While our current study has provided valuable insights into air quality prediction using the MoreAir dataset, we are actively planning and working on several key directions for future research. These directions encompass maintaining continuous data collection and real-time monitoring, ensuring the uninterrupted flow of critical data to enhance our research and its broader applications. We also aim to contribute to the advancement of the field of air quality forecasting by enhancing fault tolerance using federated learning techniques. Additionally, we anticipate the opportunity to develop and implement novel algorithms for air quality prediction, leveraging the accumulation of more data to explore advanced machine learning techniques and refine existing models, ultimately leading to more accurate and reliable forecasts.

DECLARATION OF COMPETING INTEREST

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

ACKNOWLEDGMENT

(Safaa Berkani and Ihsane Gryech contributed equally to this work.)

REFERENCES

[1] World Health Organization. (Sep. 22, 2021). *Air Pollution is One of the Biggest Environmental Threats to Human Health, Alongside Climate Change*. [Online]. Available: https://www.who.int/news/item/22-09-2021-new-who-global-air-quality-guidelines-aim-to-save-millions-of-lives-from-air-pollution

[2] *State of Global Air 2020. Special Report*, Health Effects Institute, Boston, MA, USA, 2020.

[3] R. Fuller et al., "Pollution and health: A progress update," *Lancet Planet. Health*, vol. 6, no. 6, pp. e535–e547, 2022.

[4] Elaine Ruth Fletcher. (2022). *Health Policy Watch—Health and Environment*. [Online]. Available: https://healthpolicy-watch.news/africa-faces-1-million-deaths-annually-from-air-pollution-second-only-to-malnutrition/

[5] A. Farrow, K. A. Miller, and L. Myllyvirta, *Toxic Air: The Price of Fossil Fuels*. Seoul, South Korea: Greenpeace Southeast Asia, 2020.

[6] O. Kracht, J. L. Santiago, and F. Martin, "Spatial representativeness of air quality monitoring sitesoutcomes of the FAIRMODE/AQUILA intercomparison exercise," JRC, Publications Office Eur. Union, Luxembourg, Tech. Rep. EUR 28987 EN, JRC108791, 2018.

[7] Y. Hu, G. Dai, J. Fan, Y. Wu, and H. Zhang, "BlueAer: A fine-grained urban $PM_{2.5}$ 3D monitoring system using mobile sensing," in *Proc. 35th Annu. IEEE Int. Conf. Comput. Commun.*, Apr. 2016, pp. 1–9.

[8] R. Zalakeviciute, M. Bastidas, A. Buenaño, and Y. Rybarczyk, "A traffic-based method to predict and map urban air quality," *Appl. Sci.*, vol. 10, no. 6, p. 2035, Mar. 2020.

[9] A. Wang, J. Xu, R. Tu, M. Saleh, and M. Hatzopoulou, "Potential of machine learning for prediction of traffic related air pollution," *Transp. Res. D, Transp. Environ.*, vol. 88, Nov. 2020, Art. no. 102599.

[10] L. Weissert, K. Alberti, E. Miles, G. Miskell, B. Feenstra, G. S. Henshaw, V. Papapostolou, H. Patel, A. Polidori, J. A. Salmond, and D. E. Williams, "Low-cost sensor networks and land-use regression: Interpolating nitrogen dioxide concentration at high temporal and spatial resolution in southern California," *Atmos. Environ.*, vol. 223, Feb. 2020, Art. no. 117287.

[11] B. Zhao, L. Yu, C. Wang, C. Shuai, J. Zhu, S. Qu, M. Taiebat, and M. Xu, "Urban air pollution mapping using fleet vehicles as mobile monitors and machine learning," *Environ. Sci. Technol.*, vol. 55, no. 8, pp. 5579–5588, Apr. 2021.

[12] I. Gryech, Y. Ben-Aboud, B. Guermah, N. Sbihi, M. Ghogho, and A. Kobbane, "MoreAir: A low-cost urban air pollution monitoring system," *Sensors*, vol. 20, no. 4, p. 998, Feb. 2020.

[13] S. Berkani, B. Guermah, M. Zakroum, and M. Ghogho, "Spatio-temporal forecasting: A survey of data-driven models using exogenous data," *IEEE Access*, vol. 11, pp. 75191–75214, 2023, doi: 10.1109/ACCESS.2023.3282545.

[14] R. O. Awichi, "Spatiotemporal predictions using an MSSA approach," M.S. thesis, Dept. Appl. Statist., DISS, Johannes Kepler Univ. Linz, Linz, Austria, Tech. Rep., 2015.

[15] L. Boniardi, F. Nobile, M. Stafoggia, P. Michelozzi, and C. Ancona, "A multi-step machine learning approach to assess the impact of COVID-19 lockdown on NO_2 attributable deaths in Milan and Rome, Italy," *Environ. Health*, vol. 21, no. 1, pp. 1–14, Dec. 2022, doi: 10.1186/s12940-021-00825-9.

[16] M. Stafoggia, T. Bellander, S. Bucci, M. Davoli, K. de Hoogh, F. de'Donato, C. Gariazzo, A. Lyapustin, P. Michelozzi, M. Renzi, M. Scortichini, A. Shtein, G. Viegi, I. Kloog, and J. Schwartz, "Estimation of daily PM_{10} and $PM_{2.5}$ concentrations in Italy, 2013–2015, using a spatiotemporal land-use random-forest model," *Environ. Int.*, vol. 124, pp. 170–179, Mar. 2019, doi: 10.1016/j.envint.2019.01.016.

[17] Y. Rybarczyk and R. Zalakeviciute, "Assessing the COVID-19 impact on air quality: A machine learning approach," *Geophys. Res. Lett.*, vol. 48, no. 4, Feb. 2021, Art. no. e2020GL091202, doi: 10.1029/2020gl091202.

[18] Y. Rybarczyk and R. Zalakeviciute, "Machine learning approaches for outdoor air quality modelling: A systematic review," *Appl. Sci.*, vol. 8, no. 12, p. 2570, Dec. 2018, doi: 10.3390/app8122570.

[19] I. Gryech, M. Ghogho, H. Elhammouti, N. Sbihi, and A. Kobbane, "Machine learning for air quality prediction using meteorological and traffic related features," *J. Ambient Intell. Smart Environ.*, vol. 12, no. 5, pp. 379–391, Sep. 2020.

[20] A. Masood and K. Ahmad, "A review on emerging artificial intelligence (AI) techniques for air pollution forecasting: Fundamentals, application and performance," *J. Cleaner Prod.*, vol. 322, Nov. 2021, Art. no. 129072.

[21] G. K. Uyanık and N. Güler, "A study on multiple linear regression analysis," *Proc.-Social Behav. Sci.*, vol. 106, pp. 234–240, Dec. 2013, doi: 10.1016/j.sbspro.2013.12.027.

[22] N. Zaini, L. W. Ean, A. N. Ahmed, and M. A. Malek, "A systematic literature review of deep learning neural network for time series air quality forecasting," *Environ. Sci. Pollut. Res.*, vol. 29, no. 4, pp. 4958–4990, Jan. 2022, doi: 10.1007/s11356-021-17442-1.

[23] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[24] G. I. Drewil and R. J. Al-Bahadili, "Air pollution prediction using LSTM deep learning and metaheuristics algorithms," *Meas., Sensors*, vol. 24, Dec. 2022, Art. no. 100546, doi: 10.1016/j.measen.2022.100546.

[25] G. Zheng, W. K. Chai, J.-L. Duanmu, and V. Katos, "Hybrid deep learning models for traffic prediction in large-scale road networks," *Inf. Fusion*, vol. 92, pp. 93–114, Apr. 2023, doi: 10.1016/j.inffus.2022.11.019.

[26] X. Ouyang, Y. Yang, Y. Zhang, and W. Zhou, "Spatial–temporal dynamic graph convolution neural network for air quality prediction," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Shenzhen, China, Jul. 2021, pp. 1–8, doi: 10.1109/IJCNN52387.2021.9534167.

[27] V.-D. Le, "Spatiotemporal graph convolutional recurrent neural network model for citywide air pollution forecasting," 2023, *arXiv:2304.12630*.

[28] D. Iskandaryan, F. Ramos, and S. Trilles, "Graph neural network for air quality prediction: A case study in Madrid," *IEEE Access*, vol. 11, pp. 2729–2742, 2023, doi: 10.1109/ACCESS.2023.3234214.

[29] Y. Lin, N. Mago, Y. Gao, Y. Li, Y.-Y. Chiang, C. Shahabi, and J. L. Ambite, "Exploiting spatiotemporal patterns for accurate air quality forecasting using deep learning," in *Proc. 26th ACM SIGSPATIAL Int. Conf. Adv. Geograph. Inf. Syst.*, Nov. 2018, pp. 359–368.

[30] H. Lira, L. Martí, and N. Sanchez-Pi, "A graph neural network with spatio-temporal attention for multi-sources time series data: An application to frost forecast," *Sensors*, vol. 22, no. 4, p. 1486, Feb. 2022, doi: 10.3390/s22041486.

[31] I. Gryech, M. Ghogho, C. Mahraoui, and A. Kobbane, "An exploration of features impacting respiratory diseases in urban areas," *Int. J. Environ. Res. Public Health*, vol. 19, no. 5, p. 3095, Mar. 2022.

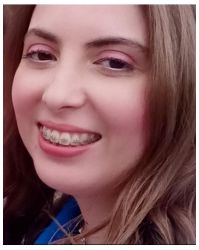
[32] *Moreair*. Accessed: Oct. 12, 2023. [Online]. Available: http://www.moreair.info

[33] *ImputeTS: Time Series Missing Value Imputation*. Accessed: Dec. 6, 2022. [Online]. Available: https://cran.r-project.org/web/packages/imputeTS/index.html

SAFAA BERKANI received the B.Sc. degree in fundamental mathematics and the M.Sc. degree in applied analysis and statistics engineering from Abdelmalek Essaadi University, Tetouan, in 2018 and 2020, respectively. She is currently pursuing the Ph.D. degree in data science with the International University of Rabat (UIR), Morocco. In 2021, she received a research grant to join the TICLab and work on her Ph.D. study. She became a temporary Lecturer with ESIN, UIR. Her research interests include data science, machine learning, and forecasting.

IHSANE GRYECH received the bachelor's degree in mathematics and informatics from Mohammed V University in Rabat, the master's degree in big data from the International University of Rabat, in 2017, and the Ph.D. degree in AI for air quality prediction and computer science from the National School for Computer Science (ENSIAS), Mohammed V University in Rabat, and the International University of Rabat, in 2023. She is a Postdoctoral Researcher with the WaveCore group, Department of Electrical Engineering, Katholieke Universiteit Leuven, Belgium. She joined the TICLab, as a Trainee, where she worked on web content by leveraging user intuition to predict items' popularity in social networks and sentiment analysis using social media. In 2019, she received the Google Africa Ph.D. Fellowship. Her research interests focus on AI for sustainable development and the application of machine learning to the environment and global health.

MOUNIR GHOGHO (Fellow, IEEE) received the M.Sc. and Ph.D. degrees from the National Polytechnic Institute of Toulouse, France, in 1993 and 1997, respectively. He was an EPSRC Research Fellow with the University of Strathclyde, Scotland, U.K., from September 1997 to November 2001. In December 2001, he joined the School of Electronic and Electrical Engineering, University of Leeds, England, U.K., where he was promoted to a Full Professor, in 2008. While still affiliated with the University of Leeds, in 2010, he joined the International University of Rabat, where he is currently the Dean of the College of Doctoral Studies and the Director of TICLab (ICT Research Laboratory). He has coordinated around 20 research projects and supervised over 35 Ph.D. students in the U.K. and Morocco. His research interests are in machine learning, signal processing and wireless communication. He is a fellow of the Asia–Pacific AI Association (AAIA). He was a recipient of the 2013 IBM Faculty Award and the 2000 U.K. Royal Academy of Engineering Research Fellowship. He is the Co-Founder and the Co-Director of the CNRS-Associated International Research Laboratory DataNet, in the field of big data and artificial intelligence. He served as an Associate Editor for many journals, including *IEEE Signal Processing Magazine* and IEEE TRANSACTIONS ON SIGNAL PROCESSING.



BASSMA GUERMAH received the Engineer degree in software engineering from the National Institute of Statistics and Applied Economics (INSEA), in 2014, and the Ph.D. degree in computer science and telecommunications from the National Institute of Posts and Telecommunications (INPT), in 2018. She is an Assistant Professor with the Computer Science Engineering School, International University of Rabat (UIR), and a member of TICLab. Her research activities revolve around machine learning/deep learning (artificial intelligence), signal processing, robotics, context-aware service-oriented computing, ontologies, and semantic web.

ABDELLATIF KOBBANE (Senior Member, IEEE) received the M.S. (research) degree in computer science, telecommunication and multimedia from Mohammed V-Agdal University, Morocco, in 2003, and the Ph.D. degree in computer science from Mohammed V-Agdal University and the University of Avignon, France, in September 2008. He has been a Full Professor with Ècole Nationale Superieure dInformatique et d'Analyse des Systemes (ENSIAS), Mohammed V University in Rabat, Morocco, since 2009. He is an Adjunct Professor with the L2TI Laboratory, Paris 13 University, France. He has more than a ten years of computer sciences and telecom experience, in Europe (France) and in Morocco, in the areas of performances evaluation in wireless mobile networks, mobile cloud networking, cognitive radio, ad-hoc networks, and future network 5G. He is the author of several scientific publications in top IEEE conferences and journals, including IEEE ICC, IEEE Globecom, IWCMC, ICNC, and IEEE WCNC. His research interests include with the field of wireless networking, performance evaluation using advanced technique in game theory and MDP in wireless mobile networks, the IoT, SDN, NFV, 5G networks, resources management in wireless mobile networks, cognitive radio, mobile computing, mobile social networks, caching and backhaul problem, beyond 5G, and future networks. He is a Senior Member of ComSoc IEEE, an Ex-Secretary of ExCom IEEE Morocco Section, the Vice Chair of IEEE Communication Software Technical Committee, and an Ex-President and a Founder of the Association of Research in Mobile Wireless Networks and Embedded Systems (MobiTic), Morocco. He is the TPC Co-Chair of IEEE ICC 2020, the TPC Chair of Wireless Networking Symposium of the International Wireless Communications Mobile Computing Conference (IWCMC 2019), the General Co-Chair of WINCOM 2020 and 2015, and the Executive Chair of WINCOM 2017. He is a responsible of master's (M.S.) of Internet of Thing and Mobiles Services (IOSM).