IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 30 January 2023, accepted 20 February 2023, date of publication 2 March 2023, date of current version 24 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3251346

RESEARCH ARTICLE

A Complete Air Pollution Monitoring and Prediction Framework

JOVAN KALAJDJIESKI, KIRE TRIVODALIEV, GEORGINA MIRCEVA, SLOBODAN KALAJDZISKI, AND SONJA GIEVSKA

Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, 1000 Skopje, North Macedonia

Corresponding author: Jovan Kalajdjieski (jovan.kalajdzhieski@finki.ukim.mk)

ABSTRACT The issue of air pollution is increasingly prominent and represents a significant environmental challenge, particularly in urban areas affected by rising migration rates. Air pollution forecasting is crucial for understanding the mechanisms underlying pollution in a specific region, but analyzing high-dimensional data with spatial and temporal dependencies poses a major challenge for traditional machine learning approaches. Additionally, missing sensor measurements due to malfunctions and connectivity loss have severely limited air pollution forecasting models' performance and restricted their use in production systems. Although significant efforts have been made in air pollution forecasting, many approaches face challenges in dealing with missing sensor data. Based on past and current research, this paper proposes and evaluates four encoder-decoder architectures with attention for forecasting particulate matter (PM) levels that are location- and season-independent. To handle missing sensor data, this paper also proposes and evaluates two adversarial networks for data augmentation. We conducted experiments to investigate the performance of predictive models with and without augmenting training datasets, and using the proposed adversarial models for data augmentation resulted in superior performance gains. The deep neural architectures developed in this research are general enough for predictive and generative tasks for other pollutants and can be adapted for handling time series data in other domains.

INDEX TERMS Adversarial data augmentation, attention adversarial data augmentation, air pollution monitoring, air pollution prediction, attention air pollution prediction, augmenting sensor data, deep learning, generative adversarial networks, recurrent generative adversarial networks.

I. INTRODUCTION

The recurring problem of air pollution is highlighted as a single largest environmental health risk in Europe [eea.europa.eu] and across the globe. Raising awareness of dangerous effects air pollution has on people's health, climate and our environment are at the forefront of government's and various stakeholders agendas. From our large-scale air pollution monitoring and forecasting system that takes a holistic view on this complex problem [1], to the models for multimodal forecasting we have proposed [2], [3], efforts to tackle the problem of air pollution has been the central objective of our research efforts in the past years. Learning representations of complex high-dimensional data generated from IoT devices or real-time streaming data that is sequential in nature and manifest characteristics such as trends and seasonality is challenging. Harnessing the power of machine learning for forecasting air pollution has received heightened attention in the last decade. The forecasting models could take on the task of predicting the air pollutant level [4], [5], [6], [7], the air pollution index [8], [9], [10], [11] or that time-to-event prediction [12]. The pollutants frequently targeted by research studies on air pollution forecasting include particle pollutants (i.e. $PM_{2.5}$ and PM_{10}) or greenhouse gases (i.e. NO_2, CO, O_3 and SO_2). A variety of auxiliary information, such as: meteorological data (i.e. temperature, humidity, wind speed and rainfall) or weather forecast data [8], [9], [13], [14] are commonly included to condition the prediction.

This study reports on several predictive and generative models for air pollution forecasting and data augmentation

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

that are based on state-of-the-art deep learning architectures to address both tasks. The levels of air pollutants including $PM_{2.5}$ concentration vary across areas and time, hence the problem we face is forecasting multi-variate time series with a spatial dimension. The notion of time and location have been the primary motivation for taking a general modelling approach with an ultimate objective to propose predictive and generative models that have applicability that is both, location- and time-independent. We have experimented with several encoder-decoder architectures with attention for $PM_{2.5}$ forecasting. The use of the attention mechanism to give a selective focus on the elements in the multivariate sequence was crucial for the performance gains yielded on the prediction task.

While extending on our past research efforts in air pollution forecasting, we also cast light on the underlying issues that tend to weaken the performance standing of the forecasting models [2], [13]. One of the major obstacles to accurate air pollution forecasting is related to the shortcomings of the embedded sensor technologies. Sensor malfunctions and loss in connectivity for brief or longer time periods are recurrent problems causing many sensor data from the measurement sequences to be missing. Decisions regarding data augmentation should be considered crucial for increasing the performance of a practical air pollution forecasting system in the presence of missing sensor data. Augmenting datasets with synthetic data samples that follow the realistic data distribution have been proposed to mitigate the problem of missing pollution data. Accounts along this line expanded from approaches relying on simple imputation methods [15], [16], [17], to methods exploiting the advances of diverse deep learning architectures [18], [19]. New lines of research using Generative Adversarial Networks (GANs) are being acknowledged for their potential to learn data in an unsupervised manner [20], [21]. Our work falls under this category, namely we introduce two novel adversarial architectures that generate realistic samples subjected to a conditional input, which have led to superior performance results.

We summarize the key contributions as follows:

- Encoder-decoder architecture with attention is proposed as a general forecasting model that is conditioned on historic air pollutant observations and a variety of auxiliary information.
- Two adversarial networks are put forward in this research to alleviate the problem of missing sensor data. To better capture the trends in the multivariate time series data, the generation of realistic samples was conditioned on a number of spatiotemporal and meteorological information.
- Of principle interest to our research are the effects of using attention-based variants for both, the predictive and generative models. To the best of our knowledge, encoder-decoder with attention architecture has not been thoroughly investigated for both task in any domain, including air pollution.

At the onset of our paper, we present a summary of current research trends in the field with a focus on studies closely related to our objectives. Section III gives the details on the creation of a dataset pertaining to real time $PM_{2.5}$ observations for the city of Skopje. An overview description of the four forecasting models advanced in our research is presented in Section IV that also elaborates on the advantages and limitations of the predictive models. The adversarial networks we have proposed for account for the missing pollutant measurements are presented in Section V. The findings of the extensive empirical experiments carried out to evaluate the proposed predictive and generative models are discussed in Section VI and Section VII, respectively. Section VIII gives concluding remarks on the relevance and contributions of this research and points to potential avenues for further research.

II. RELATED RESEARCH

Air pollution forecasting is a worthwhile endeavor and the increasing attention it has attracted coincides with the recent advances in deep neural networks. As with past imports of deep learning in other application domains, the use of deep neural networks carries both great promise and many challenges that still await robust and general solutions. Air pollution forecasting models vary in their architectures, although Convolutional Neural Networks (CNNs) [22] and Recurrent Neural Networks (RNNs) [23] are the most popular approaches.

Recurrent neural networks (RNNs) [6] have been a staple approach for modeling multivariate time series data inherent to air pollution forecasting advanced in recent research. Recurrent neural network variants, Long Short-Term Memory (LSTM) networks [10] and Gated Recurrent Unit (GRU) are the most frequently used architectures for modeling complex spatial, temporal relationships in multivariate sequence data. Weibul-time-to-event Recurrent Neural Network (WTTE-RNN) [12], autoencoder model [5], [24], convolutional neural networks [25], [26], [27], as well as ensembles methods [9] have also been explored for air pollution forecasting.

In the last few years, air pollution prediction models are augmented with attention to determine which element or variable in the multivariate data should be given importance [13], [14]. In [13], an LSTM model with attention has been used to determine the importance of wind direction and speed as part of a prediction model based on an ensemble learning method. Cheng et al. [14] have proposed a generic neural model, using feed-forward and recursive neural networks, that employs attention to dynamically assign an importance degree to the monitoring stations as a basis for prediction in an area without monitoring sensors. In contrast to the joint modeling of the observational and auxiliary data, our predictive models learn the complex dependencies of the multivariate input data by using encoder-decoder architecture with attention.

Many efforts have been taken to address the problem of missing sensor values in air pollution forecasting, from a

simple method of removing all samples with missing measurements to statistical imputation to the newest GAN-based approaches. The early practices of simple imputation methods were replaced by data augmentation methods based on deep learning and today, quite wide differences of architectures are proposed for data augmentation of sequential data in different contexts and domains.

A large body of research on augmenting time series data has focused on applying LSTM GAN-based architectures [28], [29], [30]. Convolutional networks [22] have also been used as building blocks of generative adversarial networks [31], [32]. The performance advantages of data augmented training were noted for the time-conditional T-CGAN model proposed in [31], especially in cases of small datasets with noisy and irregular data. We have selected recent research studies closely related to our proposed models. An LSTM adversarial model, employed for the problem of data augmentation of biosignals (e.g. EEG and ECG signals), has obtained notable performance gains on a related classification task [28]. A Recurrent GAN (RGAN) for generating sequences of medical real-valued time series data was presented in [30]. The paper also introduces a novel method that has become a standard approach for evaluating the performance of GAN-based architectures named ''Train on Synthetic, Test on Real'' (TSTR). We have followed the same approach when testing our models. The authors have also proposed a Recurrent Conditional GAN (RCGAN), as a variant suitable to generate realistic data subject to some conditional input, namely labeled training data samples. While conditioning the generation process was of special importance in our data augmentation models, the generation of new synthetic samples was controlled by a larger set of spatiotemporal and auxiliary i.e. meteorological information.

An adversarial model TimeGAN [33], closely related to our approach, simultaneously learns to encode features, generate representations, and iterate across time. It includes two additional autoencoding elements, called an embedding and recovery networks, that are trained jointly with the adversarial components, the generator and the discriminator. The embedding network provides the latent space, the adversarial network operates within this space, and the latent dynamics of both real and synthetic data are synchronized through a supervised loss.

GAN architectures using a modified Gate Recurrent Unit (GRU) have been used for imputation of multivariate time series data [34], to handle both, temporal irregularities and missing data. The proposed model was evaluated on two datasets, including the air pollution KDD CUP 2018 Dataset [35], highlighting the superiority on the prediction and classification tasks when imputed datasets were used.

Convolutional networks [22] have also been used as building blocks of generative adversarial networks [31], [32]. The comparative analysis of the proposed models evaluated on a number of benchmark and publicly available synthetic and real-world datasets The performance advantages of data augmented training were noted for the time-conditional T-CGAN model proposed in [31], especially in cases of small datasets with noisy and irregular data.

III. DATA

Investigative studies have confirmed that pollutants, such as particulate matter (PM), which refers to a mixture of solid and liquid particles present in the air are highly correlated with a high incidence of cardiovascular and lung diseases. To this end, a case study on forecasting $PM_{2.5}$ pollution for the city of Skopje, frequently listed as one of the most polluted cities in the world[1] was carried out to train and evaluate the predictive and generative models proposed in this research.

In line with our long-term objective to design and deploy an IoT holistic system for monitoring and forecasting air pollution in urban areas [1], [2], [3], [36], we have opted for a real-world $PM_{2.5}$ pollution dataset, created for the city of Skopje that was in the focus of our case study. For the purpose of evaluating the models for forecasting $PM_{2.5}$ pollution levels and the generative models for data augmentation, we have obtained 865,453 $PM_{2.5}$ measurements for the period extending from April 2017 to March 2020 available from pulse.eco.[2]

There are 40 monitoring stations, distributed across the city area as shown in Figure 1. The monitoring stations are equipped to sense various sets of pollutants, although $PM_{2.5}$ concentration was chosen to be targeted in our study, since the particle pollutant is measured at all 40 locations. The time step at which the PM2.5 values are measured and transmitted to the website vary across the measuring stations. To account for the irregularly-sampled data, the measured $PM_{2.5}$ levels were aggregated and the mean of the $PM_{2.5}$ concentration measured for each hour was calculated to represent the hourly $PM_{2.5}$ concentrations. Each $PM_{2.5}$ measurement was assigned the corresponding timestamp (date and time) and the location i.e., the latitude and longitude of the measuring station.

Seasonality and trends are inherent to variations in air pollutant concentrations. The weather conditions are known to affect air pollution, especially in the case of the city of Skopje that sprawls across a valley surrounded by high mountains. The meteorological variables, considered in our study as auxiliary features, were obtained from the API endpoints of World Weather Online[3] on an hourly basis. They include: weather description, temperature, humidity, pressure, dew point, precipitation, visibility, cloud coverage, heat index, UV index, wind direction, wind speed, wind chill and wind gusts. Among these features, the categorical variables, such as weather description, wind direction, heat index and UV index were represented as one-hot encodings, while the other numerical variables were normalized in the range of [0, 1].

[1] https://www.matchup-project.eu/news/air-pollution-in-skopje-how-citizens-spurred-policymakers-towards-the-change/ last accessed: 15.01.2023.

[2] https://pulse.eco/ last accessed: 15.01.2023.

[3] https://www.worldweatheronline.com/ last accessed 15.01.2023.

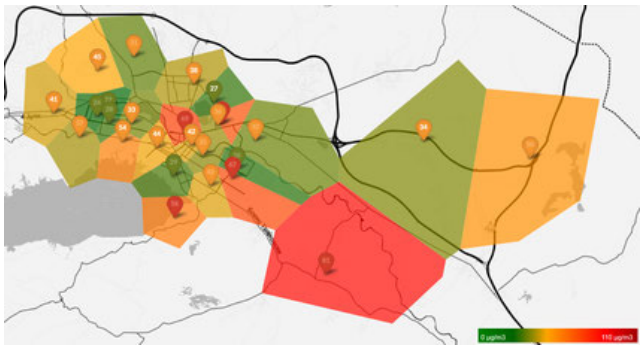

FIGURE 1. Location of current active sensors on the map of Skopje. The coloring map is shown below, and measures the $PM_{2.5}$ pollution. The map is taken from https://skopje.pulse.eco/.

TABLE 1. Missing measurements across sensors and pollutants on a yearly basis.

Year	Missing Values	Percentage Missing Values
2017	14688	16.81%
2018	56359	16.08%
2019	65303	18.63%
2020	50056	19.17%

FIGURE 2. Hourly missing values across sensors and across pollutants.

Current measuring platform is notorious for a large number of missing measurements due to various problems associated with the currently employed sensor technologies. The distribution of the missing values aggregated across all monitoring stations and across all measured pollutants on hourly and yearly basis are presented in Table 1 and Figure 2, respectively. It should be noted that even though, the percentages of missing values are in the range of 15% to 20% percent across various pollutants, the $PM_{2.5}$ sensors did not suffer from such a serious downtime periods. Moreover, we have safeguarded the training process against sequences that have more than two consecutive missing values.

IV. AIR POLLUTION FORECASTING MODELS

The models for forecasting pollutant levels advanced in our research are based on the Encoder-Decoder framework, proposed by [37] and [38] upon which the architecture with attention was built [39]. The performance standing of several variants of this end-to-end architecture were evaluated on the dataset for the city of Skopje. The attention mechanism was leveraged to discriminate the importance of various auxiliary variables, such as weather conditions, known to affect the concentration levels of air pollutant.

A. PROBLEM STATEMENT

The measured $PM_{2.5}$ concentration value, the weather conditions, the latitude and longitude of the monitoring station's location, and the temporal information, i.e., the date, time and the season when the observation was taken were concatenated to form a 75-feature vector that was fed as input in the predictive aproposed in this research. The multivariate vector at the $i-th$ hour $X = \{M_i, H_i, L_i, W_i\}$ contains a selected set of static and dynamic features related to a particular $PM_{2.5}$ observation. The first element, M_i, denotes the measured $PM_{2.5}$ concentration levels at the $i-th$ hour, while H_i represents the temporal information i.e., date, the hour and the season when the $PM_{2.5}$ observation was taken. The two-feature vector L_i indicates the location of the $PM_{2.5}$ monitoring station given by its latitude and longitude. The weather conditions at each hour were represented by a multi-feature vector W_i, containing both static and dynamic data.

The task of $PM_{2.5}$ forecasting is defined as follows: for a given input sequence of observations $\{X_i\}_{i=1}^{T_x}$ the model is expected to predict the $PM_{2.5}$ values for the subsequent k hours. Previous research studies have focused on predicting air pollutant concentrations after having observed the trends in air pollution in a given time frame, usually 24 hours period. In contrast, rather than studying short time slices, we observe the trends over longer time periods, 7 to 14 days, to draw better inferences about the complex dependencies at play. For the case study under investigation, a history window length T_x was set to 7 days (168 hours) and k was set to 24, which means that based on the historic observations in the last 7 days, the $PM_{2.5}$ concentration values for the next 24 hours were predicted.

B. OVERVIEW OF LONG SHORT- TERM MEMORY (LSTM) NETWORK

Long Short-Term Memory (LSTM) networks [40] as special types of recurrent neural networks, have been a staple approach for modeling multivariate time series data. Our encoder-decoder models employ the Long Short-Term Memory (LSTM) to capture the long-term dependencies inherent to air pollution data. An LSTM leverages the gating mechanism to control the information between the input, output and the cell memory. At each time step, the equations describing how an LSTM cell works follow:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \quad (1)$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \quad (2)$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \quad (3)$$

$$C_t = f_t \odot C_{t-1} + i_t \odot tanh(W_C x_t + b_C) \quad (4)$$

$$h_t = o_t \odot tanh(C_t) \quad (5)$$

where i_t, f_t, and o_t denote the input gate, forget gate, and output gate that control what is being stored, deleted or outputted by the cell state C_t. For a given input vector x_t at time t, the gating vectors (1) to (3) determine the historic information that is kept in the cell memory. The cell state C_t is updated according to equation (4), while equation 5 describes the update of the current hidden state denoted as h_t. The sigmoid and hyperbolic tangent activation functions are denoted by σ and *tanh*, respectively. In our implementations, the Rectified Linear Unit (ReLU) activation functions were used for f_t. The linear transformation matrices W and the bias vectors b are labeled with subscript indices i, f, o, and C referring to the input gate, forget gate, output gate and the cell state, respectively. Pairwise multiplication $\odot$ denotes the Hadamard product.

1) AN LSTM-BASED ENCODER-DECODER WITH ATTENTION

In this research, LSTM-based encoder-decoder architectures with attention has been investigated for the task of predicting PM2.5 concentration. The encoder-decoder architecture, originally proposed for the task of neural machine translation [39], are suitable for learning the complex dependencies in multivariate input sequence and generate an output sequence of an arbitrary length

In the forecasting models presented in this Section, the LSTM encoder network learns the representation of the multivariate 75-feature input vector, by taking the elements from the input sequence $X = \{X_1, X_2, \ldots, X_{T_x}\}$, sequentially. The cell state and the hidden state are recursively updated according to equations (1) through (5).

The LSTM decoder network iteratively outputs the predicted $PM_{2.5}$ values for the next k time steps $Y = \{y_{Tx+1}, \ldots, y_{Tx+k}\}$, one value at each time step. In other words, the predicted $PM_{2.5}$ value y_t is an approximation of the conditional probability

$$p(y_t|x_1, \ldots x_{T_x}) = \prod_{t=T_x+1}^{T_x+k} p(y_t|c_t, y_1, \ldots, y_{t-1}) \quad (6)$$

An attention mechanism is employed to discriminate the importance of the historic $PM_{2.5}$ measurements, while also leveraging auxiliary information in the input multivariate vector to account for the temporal and spatial information as well as the weather conditions at the time of the measurement. In other words, the prediction is based on the representation learned by the encoder by giving selective focus to the elements in the input time series that account more in the current prediction step.

Attention weights e_{ij}, also known as attention scores are assigned to different features of the input vector depending on its relevance for current $PM_{2.5}$ prediction. Dot-product attention has been used as follows:

$$e_{ij} = a(s_{i-1}, h_i) = [s_{i-1}^T h_1, .., s_{i-1}^T h_{T_x}] \quad (7)$$

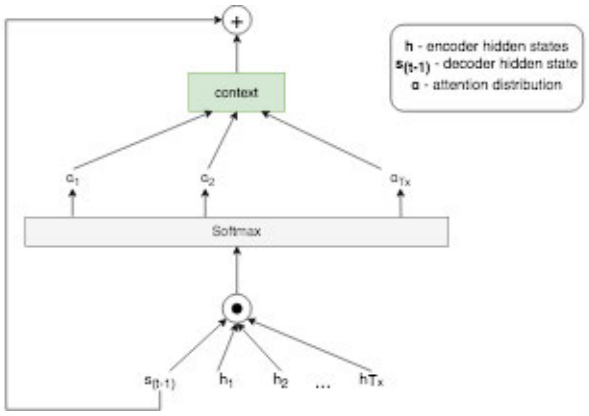

FIGURE 3. Attention block.

The attention distribution is obtained by a softmax function as follows:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})} \quad (8)$$

At each time step, the output of the attention block is an attention context vector c_i, computed as a weighted sum of the attention distribution α_{ij} over the hidden state of the encoder:

$$c_i = \sum_{j=1}^{j=T_x} \alpha_{ij} h_j \quad (9)$$

where, h_j refers to the hidden state of the encoder LSTM cells. The attention context vector c_t is concatenated with the decoder hidden state s_{t-1} and the predicted output y_{t-1} from the previous time step. The output layer acts as a regression function that outputs the $PM_{2.5}$ predicted value y_t.

C. MODEL VARIANTS

The multivariate input vectors contain historic $PM_{2.5}$ observations and the temporal, spatial and weather-related features as well. A selective attention incorporated into the encoder-decoder architecture needed additional boosting to reflect the contributions of the existing complex dependencies of the auxiliary features when predicting future $PM_{2.5}$ values. Two techniques to increase the expressiveness of the attention-based architecture have been experimented with, both on the encoder network. Namely, bidirectional and stacking were proposed to for better representational learning of the complex multivariate data.

Two baselines and four variants of the encoder-decoder architecture with attention were explored:

- **Stacked LSTM** The basic long short-term memory network [Hoh1997LSTM] is a recurrent neural network that uses LSTM as the computing unit. Unlike the encoder-decoder architecture, a deep LSTM network outputs the PM2.5 predicted value of the next time step when fed a sequence of historic observations concatenated with the spatiotemporal and weather information.

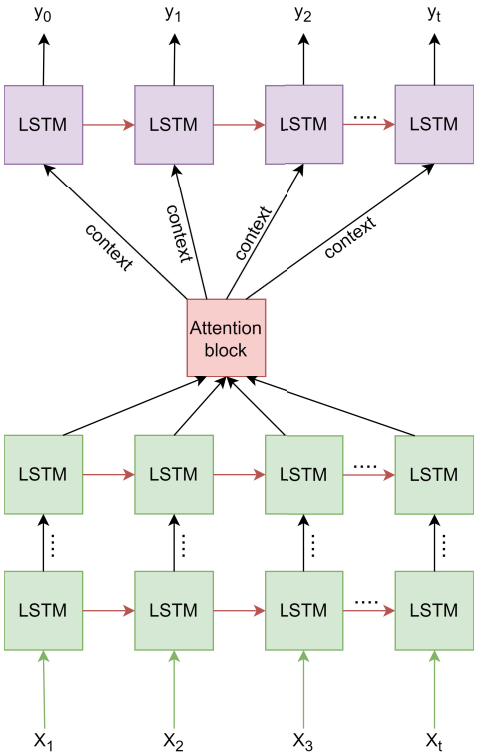

FIGURE 4. Stacked attention model architecture.

- **Bidirectional Stacked LSTM** Bidirectionality and stacking are standard techniques that can improve the representation learning of complex data. By taking the input sequence in both direction and increasing the expressiveness of the network by adding 4 stacked LSTM layers, a Bidirectional Stacked LSTM network has been created.
- **AM** - The Attention Model denotes an LSTM encoder-decoder with attention, as described in the previous subsection.
- **SAM** - The Stacked Attention Model incorporates a two-layer stacked LSTM encoder; we have experimented with various number of layers and a two-layer LSTM encoder yielded the best performance results. Stacking layers adds to the expressiveness of the model to represent the complexity of the trends and dependencies in the multivariate time series, which are learned by the model. The architecture of the SAM model is shown in Figure 4.
- **BAM** - A Bidirectional Stacked Attention Model employs a bidirectional LSTM encoder to exploit the dependencies in both direction, which is suitable for representation learning in the encoder.
- **BSAM** - A Bidirectional Stacked Attention Model leverages the advantages of both techniques; hence, a bidirectional stacked LSTM encoder was employed.

V. DATA AUGMENTATION MODELS

Air pollution forecasting models that are based on data generated by IoT sensors face the recurring problems of missing measurement data due to sensor malfunctions or loss of connectivity. In this respect, any air pollution forecasting model based on deep learning would require a component for generation of realistic data samples to augment the time series dataset.

Data augmentation techniques are viewed as a general solution to address the issues related to the quantity and quality of training datasets. The objective of equal importance posed in this research was to create models that would be able to generate synthetic data samples that reflect the realistic distribution of $PM_{2.5}$ values for the pollution area under investigation. We address the challenges of missing sensor data by proposing two models for data augmentation of time series data by utilizing state-of-the-art adversarial networks.

While drawing upon previous research on architecture variants of the Generative Adversarial Networks (GANs), introduced in the seminal work of Ian Goodfellow et al. [41], the structural changes we propose stem from the need to capture the complex dependencies in the multivariate input sequence to condition the generation process. In what follows, we describe the adversarial models we propose for data augmentation of datasets containing historic air pollutant observation.

A. GENERAL ARCHITECTURE

Generative Adversarial Networks (GANs), as originally proposed, are composed of two neural networks that are trained jointly in an adversarial manner. The Generator, G is trained to learn the real distribution of the training data and generate realistic samples that would deceive the Discriminator. When presented with a data sample, the role of the Discriminator is to determine the true origin of the sample, which might come from the training data or the generated samples. The training process has the characteristics of a minmax game, in which the Generator is trying to maximize the error of the Discriminator, by generating samples that are so realistic that the Discriminator fails to distinguish between the real and synthetic ones.

The general adversarial architecture of our generative models for data augmentation, shown in Figure 5 consists of two main modules: a Generator and a Discriminator. Formally, the Generator takes a randomly sampled noise vector Z from prior distribution p_z and maps it to a sample $G(z; \theta_g)$, where θ_g refers to the learnable parameters of the G network. The distribution of the generated samples is denoted as p_{model} and is expected to be close approximation of the real distribution p_{data} of the training samples. In our implementation, the latent noise p_z is sampled from a normal distribution with mean 0 and standard deviation 1.

$$\min \max V(G, D) = \mathbb{E}_{X \sim P_{data}(X)}[\log D(X)] + \mathbb{E}_{X, z \sim P_z(z)}[\log(1 - D(G(X, z)))] \tag{10}$$

The discriminator D is a binary classifier with parameters θ_d, which distinguishes between fake inputs generated by the Generator, and the real inputs sampled from the training

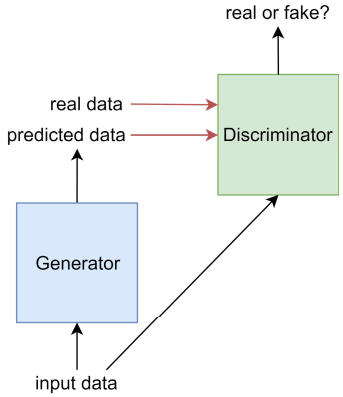

FIGURE 5. A general adversarial architecture of data augmentation models.

data p_{data}. More precisely, $D(x)$ is a scalar value, representing the probability that x is sampled form p_{data}, rather than sampled from p_{model}. The discriminator D is trained to maximize the probability of assigning a correct label to a sample drawn from the training samples and a sample generated by G. Simultaneously, the Generator G is trained to minimize log(1 - D(G(z))). In other words, D and G play the following two-player minimax game with value function $V(G, D)$. The minmax adversarial loss (maximized on the discriminator side, while at the same time minimized for the generator) ensures that the likelihood of a sample being from the training dataset or created by the generator is correctly assigned by the discriminator.

Unstable training, model collapse, and vanishing gradients are some of the major problems GAN-based models face and a lot of architectural decisions have been proposed to overcome these problems. Of principal interest in this research is to condition the generation of synthetic samples to follow the trend of the real time series data. Air pollutant concentrations change over time and are subject to variations, depending on a variety of external factors, such as weather conditions, wind and precipitation, in particular. Temporal features, such as season, date and the time of day of when the measurement was taken are highly correlated with the $PM_{2.5}$ concentration peaks as well.

The proposed architectural changes to the original GAN framework are guided by the necessity to learn the complex dependencies conveyed by the multivariate input data that should condition the generation process. In order to be able to generate a synthetic data sample that adheres to the realistic distribution of the training data, a synthetic $PM_{2.5}$ sample should be generated conditioned on the weather at the time and place it was captured. To this end, the Generator and the Discriminator are fed with a multivariate input vector that carries the same information used in the forecasting models, i.e., $X = M_i, H_i, L_i, W_i$, where M_i refers to a $PM_{2.5}$ measurement value at the $i - th$ hour, H_i and L_i are the temporal and location information and the weather conditions related to the measurement are denoted as W_i.

While adversarial networks afford training in an unsupervised manner by optimizing the minmax loss function presented in Equation (10), another loss function was added to address the particularities of the generative process. An $L2$ loss given as Equation (11) was added to guide the Generator to adhere to the distribution of the realistic samples from the training dataset. The control afforded by the supervised training operationalized through the added $L2$ loss ensures compatibility between the generated sample and the dependencies inherent to the training data.

$$\mathcal{L}_{L2}(G) = \mathbb{E}_{x,z}[\|y - G(z)\|_2] \tag{11}$$

The $L2$ loss is treated as a regularizing term and it is weighted with the hyper-parameter λ. With the L_2 loss added, the final objective function V^* becomes:

$$V^* = \arg\min\max V(G, D) + \lambda\mathcal{L}_{L_2}(G) \tag{12}$$

The rationale for the structural changes we propose to the general adversarial architecture is discussed in the next subsection.

B. MODEL VARIANTS

Two adversarial models for data augmentation were considered: one instantiating the G and D networks as deep LSTM networks and the other using an encoder-decoder architecture with attention, similar to the architectures of the $PM_{2.5}$ forecasting models we have used. These two models differ in the way the conditional representation learning is performed and in the way the discriminator classification is performed.

- **Adversarial Network Encoded (ANEncoded)** In order to be able to generate a synthetic data sample that adheres to the realistic distribution of the training data, the generation of any synthetic $PM_{2.5}$ sample should be conditioned on the factors affecting air pollutant concentration, such as the time, the place and the weather conditions associated with it. In addressing these aims, a deep LSTM network maps the distribution of the air pollutant data into a low-dimensional latent space as shown in Figure 6. The last LSTM cell of the generator is fed with an input sampled from the normal distribution Z and outputs a sample that is subjected to a conditional input i.e., the spatiotemporal and auxiliary features. The Discriminator has the role of a classifier that based on the representation learned from historic observation during training, tries to better discriminate between a sample drawn from the training or generated samples.
- **Attention Adversarial Network (AAN)** In the proposed AAN model, both the generator and discriminator are designed with deep encoder-decoder architectures as their backbones, as shown in Figure 7. The unfolded LSTM networks, employed for the generator and discriminator networks, selectively propagate information through a gating mechanism, capturing the interdependencies in the multivariate input vector that carries relevant information. To this purpose, both the generator and the discriminator networks, are fed with the

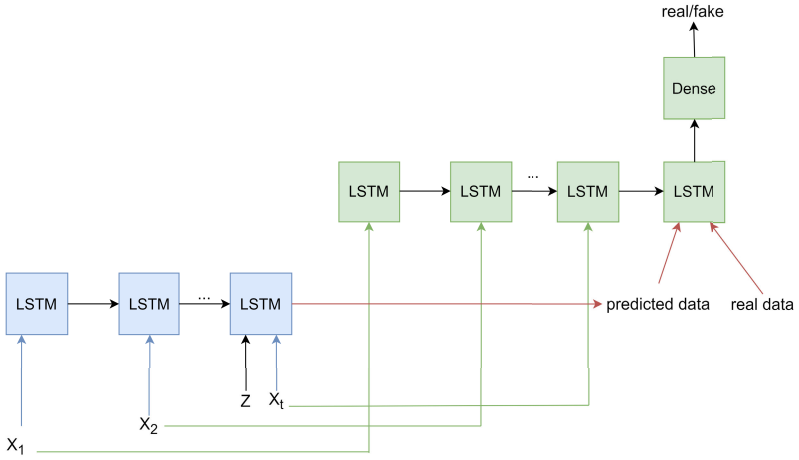

FIGURE 6. Adversarial network encoded architecture.

multivariate input vector Xi = Mi, Hi, Li, Wi so that the generation of a synthetic sample is conditioned on the historic measurements and the spatiotemporal and weather information. The attention block is employed to highlight relevant input features, while suppressing redundant information propagation through the network. In this way, the encoder-decoder generator learns the distribution of the training data, so that more realistic variations of $PM_{2.5}$ samples are generated.

VI. EVALUATION OF $PM_{2.5}$ PREDICTION MODELS

The detailed results of the performance analyses of the PM2.5 forecasting models we have proposed are discussed in this section. We have used two baseline models, namely Stacked LSTM and Bidirectional Stacked LSTM networks, against which the performance of the four forecasting models were compared to. Our $PM_{2.5}$ forecasting models are four variants of the attention-based encoder-decoder prediction models. We have compared the normalized root mean square error (RMSE) and normalized mean absolute error (MAE) of the forecasting models. Lower RMSE and MAE values indicate better performance of the models. The experimental results are reported for a 80% to 20% separation of the training dataset.

To alleviate the problems of missing measurements, simple imputation techniques, such as the last data sample before the sensor's downtime, or using the first measured value after the sensor has recovered were used. These simple imputation techniques are not a viable approach for longer sensor malfunction periods; they often yield values that are not relevant after long sensor inactivity. Forecasting $PM_{2.5}$ concentrations with the models serve as a testbed models for validating the suitability and the effectiveness of our two adversarial models for data augmentation.

The dimensions of the hidden state of the LSTM encoder and decoder have considerable effect on the complexity and the expressiveness of a model, such as the number of neurons per layer. Computational efficiency and low memory requirements were also considered. The tuning of other parameters was needed to provide efficient training and faster convergence of the models. In addressing these aims, Adam optimizer was selected to dynamically tune the learning rate during training, while the most popular regularization technique dropout was chosen to prevent our models from overfitting. The grid search of the batch size was conducted and set to 512. The initial learning step differs among the models. The hyper parameters of the proposed predictive neural models and the baseline methods have been selected through a grid search; for each model the selected settings leading to best performances are shown in Table 2.

The value of the hyper parameter history window length over which pollution inference is performed was chosen experimentally. History window length refers to the number of elements in the input sequence i.e., number of past observations that are used to predict the pollutant levels in the next 24 hours. We have experimented with multiple window lengths in the range of 7 to 14 days, to investigate its impact of the predictive power of our forecasting models. The learn the most informative data from the input, while suppressing the noise in the observed pollutant data.

The performance results of our four variants of the attention-based encoder-decoder prediction models as well as of the two baseline methods are provided in Tables 3. We have validated the quality of the forecasting models on a real-time dataset created for the city of Skopje in the period of 2017-2019. The performance advantage of the attention-based models has been evident on both, the RMSE results on the training and test dataset. The advantages of the attention mechanism per se were less distinctive without the use of bidirectionality and stacking in the encoder network. We have observed little difference in the RMSE values between the best baseline, Bidirectional Stacked LSTM and the basic Attention Model (4.15% vs 4.17%). Gains of up to a 30% improvement on predicting a 24-hour sequence was

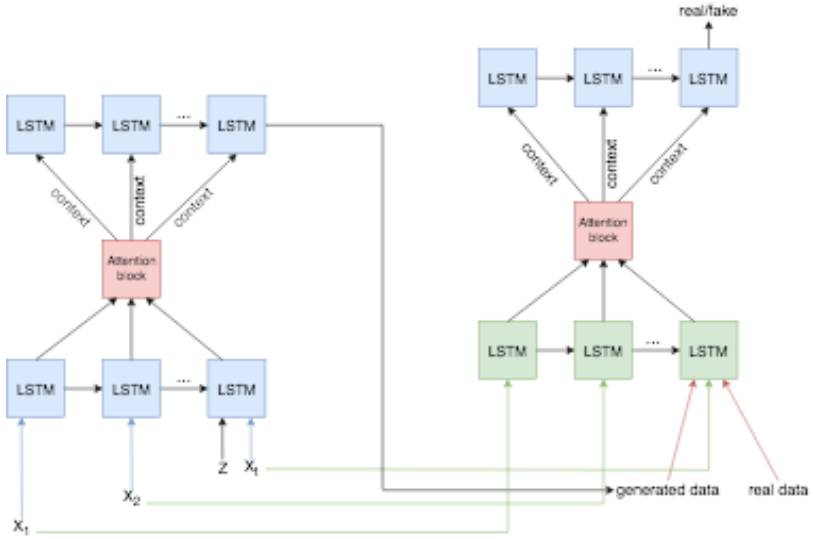

FIGURE 7. Attention adversarial network architecture.

TABLE 2. Hyper parameters of the $PM_{2.5}$ forecasting models.

Models	Stacked LSTM	AM	SAM	BAM	BSAM
Training Epochs	100	100	100	100	100
Learning rate	0.001	0.001	0.001	0.001	0.001
Clip value	1.0	1.0	1.0	1.0	1.0
Optimizer	Adam	Adam	Adam	Adam	Adam
Stacked Layers	4	/	2	/	2
Number of Cells	32, 32, 64, 128	32, 64	32, 32, 64	32, 64	32, 32, 64
Batch Size	512	512	512	512	512
Activation	ReLU	ReLU	ReLU	ReLU	ReLU

TABLE 3. $PM_{2.5}$ prediction results of our forecasting models and two baseline methods on a real-time data for the city of Skopje in the period of 2017-2019.

Models	Training MSE	Testing MSE
Stacked LSTM	4.91	8.92
Bidirectional Stacked LSTM	4.15	7.22
Attention Model	**4.17**	**7.13**
Stacked Attention Model	**3.65**	**4.57**
Bidirectional Attention Model	**3.73**	**5.52**
Bidirectional Stacked Attention Model	**1.98**	**3.41**

noted when bidirectional and stacked encoder was used in the SAM and BAM model, respectively. In contrast, the Bidirectional Stacked Attention Model (BSAM) that implements both techniques shows twice lower RMSE than the baselines. Namely, the best performing model BSAM exhibited root mean square error 1.98% compared to the baselines, Stacked LSTM and Bidirectional Stacked LSTM showing RMSE 4.91% and 4.15%, respectively). We may conclude, that the benefit of incorporating attention to better learn the trends was more prominent when the complexity of the multivariate input data was captured by a deep neural networks with matching complexity. The attention-based encoder-decoder framework shows better forecasting performance during testing as well; it outperforms the baseline models over every variant network. The proposed attention-based forecasting models had lower RMSE ranging from 7.13% for the Attention Model down to 3.41% for the BSAM during testing. The contribution of the bidirectional and stacked encoder follow similar pattern as the behavior observed during training., making the best performing model the BSAM that incorporates bidirectional stacked variant. The best performing model BSAM reduces the root mean squared error by 52%

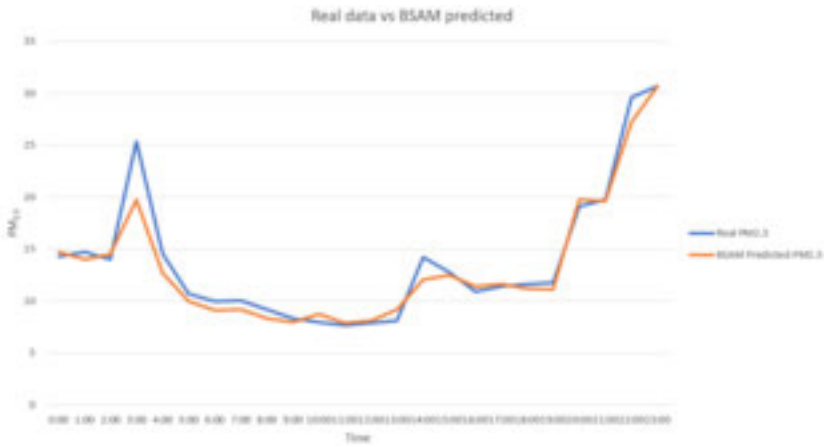

FIGURE 8. Comparison of $PM_{2.5}$ actual observations and values predicted by the best performing Bidirectional Stacked Attention Model (BSAM).

on our $PM_{2.5}$ pollution dataset compared to the best baseline method. In summary, the performance gains of the four models compared to the baseline models have proved the predictive power of the encoder-decoder with attention on the task at hand.

In an attempt to understand how the prediction unfolds, a sample of 24-hour sequence of actual $PM_{2.5}$ values was plotted along the PM2.5 values predicted by the BSAM model as shown in Figure 8. The capability of the model to learn the seasonality and trends during a 24-hour period is evident. We may infer that bidirectionality and stacking can significantly boost performance of the attention-based encoder-decoder architecture on the forecasting task. Our insights also affirm that the BSAM model captures the multivariate context well by representing the features identified as important from the history window that spans 7 days. Predicting a sequence of spatiotemporal data was performed by a powerful generative encoder-decoder model that learns from data it was exposed to. We may speculate that peak values could be related to spurs sources of pollution (e.g., industry combustion), which might require external factors and auxiliary data to be incorporated.

Figures 9 and 10 show the convergence curves of our four models and two baseline methods during training and testing, respectively. While the RMSE between-epochs changes are very small after the 20th epoch over most of the models during training, the BSAM model continues to converge and exhibits superior performance compared to others. Longer training times were expected for the BSAM model; they could be attributed to the complexity of the two-layer stacked bidirectional encoder network. During testing, the baseline models converge faster with the Bidirectional Stacked LSTM model encountering few local minimums during convergence. The convergence for the SAM, BAM and BSAM happens after the 35th epoch; among them the best performing BSAM continues for few more epochs.

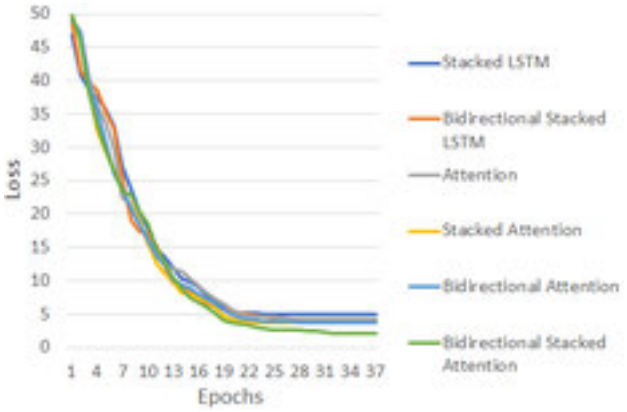

FIGURE 9. Training mean squared error for the $PM_{2.5}$ forecasting models.

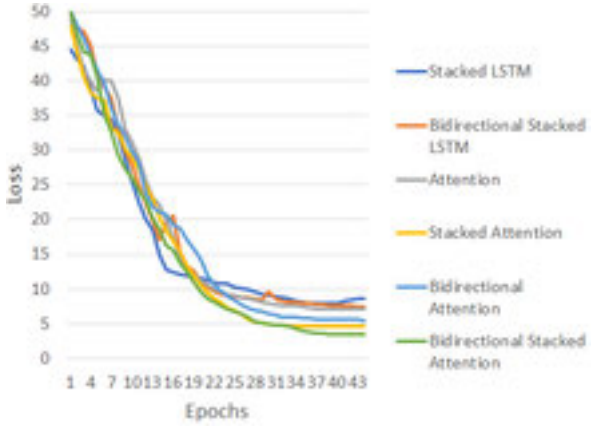

FIGURE 10. Testing mean squared error for the $PM_{2.5}$ forecasting models.

VII. EVALUATION OF DATA AUGMENTATION MODELS

A. EXPERIMENTAL SETUP

Forecasting $PM_{2.5}$ concentrations with the models serve as a testbed models for validating the suitability and the effectiveness of our two adversarial models for data augmen-

tation. The evaluation of the forecasting models discussed in previous section has been done without data augmentation of the $PM_{2.5}$ dataset used in this study. The following discussion sheds light on the added value of augmenting the training datasets with synthetic samples generated by the proposed adversarial models, ANEncoded and AAN. To address the question of whether or not, and to what extent the proposed data augmentation models contribute to increasing the performance, we needed to investigate their effectiveness on the $PM_{2.5}$ forecasting task with and without data augmentation by using: 1) a simple deep LSTM architecture using the TSTR evaluation approach and 2) our best performing forecasting model BSAM. The experimental setup of using a 48-hour historic measurements to generate a 24-step sequence of synthetic data samples was followed. The two adversarial models for data augmentation, ANEncoded and AAN were trained on a dataset that was filtered from sequences containing missing elements/measures, resulting in a 10% reduction of the originally created dataset.

The hyper parameter settings of our two adversarial architectures for data are given in Table 4. These values were chosen experimentally on a validation dataset. Grid search was employed to find the best values. We were also guided by our experience on training the forecasting models discussed in previous section.

B. RESULTS

The forecasting models proposed in the previous section serve as a testbed task for validating the suitability and the effectiveness of the proposed data augmentation models.

For evaluating the hypothesized performance advantage of augmenting dataset with samples generated by our adversarial architectures, the Train on Synthetic - Test on Real (TSTR) [refs] evaluation approach was used. A simple three-layer LSTM network was used as a generic network evaluated on the task of PM2.5 prediction in three different scenarios. In Table 5, we provide the RMSE scores exhibited by the simple three-layer LSTM network over various types of training scenarios i.e., trained for 50 epochs with and without augmented dataset. First, the three-layer LSTM model was trained on the dataset without any data augmentation; simple imputation technique was employed i.e., a missing $PM_{2.5}$ measurement was substituted with the last observation. In the other two scenarios, it was trained with a dataset augmented with synthetic data samples generated by the ANEncoded and AAN model, respectively. The simple deep LSTM network was tested on the real data samples; there were no missing elements in the input sequences. The results show that data augmentation significantly improves the model's accuracy, reducing the training loss from 8.92% obtained without data augmentation to 5.91% and 5.55%, depending on the quality of the samples generated by our adversarial networks, ANEncoded and AAN, respectively.

Adversarial networks learn the generative model from the data they are exposed to. The TSTR RMSE values during training of the simple three-layer LSTM network are depicted in Figure 11. This behavioral evidence showing stable training without oscillation is attributed to data augmentation. When it comes to handling missing data, it is evident that the simple deep LSTM forecasting model yields better performance when the training datasets are augmented with synthetic samples generated by the two adversarial networks proposed in this research.

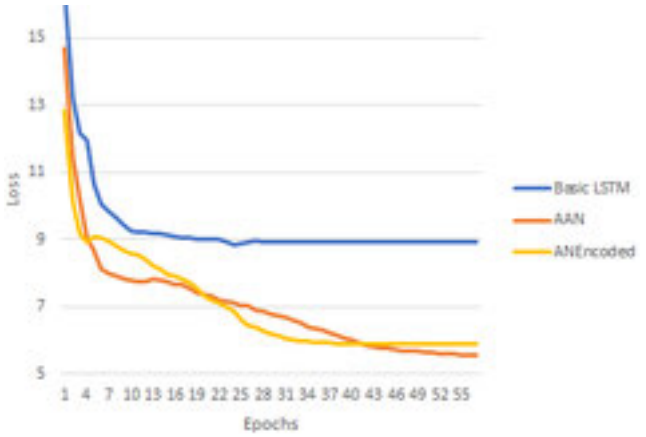

FIGURE 11. TSTR loss of the models.

Our main objective was to improve the performance of the PM2.5 forecasting models by augmenting the training dataset with synthetic data samples generated by the adversarial networks. The performance gains by data augmentation were demonstrated for our best performing PM2.5 forecasting model BSAM, which is a bidirectional stacked variant of the attention-based encoder-decoder.

The findings presented in Table 6 have showed that training the forecasting BSAM model on augmented dataset improve its performance. The performance gains ranging from 25% reduction of RMSE value when augmenting the dataset was done by the ANEncoded to 50% reduction when AAN was used for generating data samples. The improvement during testing has followed the same pattern as the one during training. During testing on the dataset containing only real data samples, the RMSE scores exhibited by BSAM has been as low as 1.27% when it was trained with data augmented by the AAN model, compared to 3.41% obtained when trained on a dataset using simple imputation to handle missing data. We may argue for the central role the encoder-decoder with attention plays in generating more realistic data. To be more specific, the encoder-decoder-structured generator learns the latent representation of the multivariate training data that condition the generation of a synthetic sample, while the encoder-decoder structured discriminator assigns a correct likelihood to a sample being drawn from the training set or being a synthetic one created by the generator. The findings showed that adding attention to selectively focus or condition the output of the models increased the performances of the generative models w/o attention as well the performance of the predictive model.

TABLE 4. Best hyperparameter values for the adversarial architectures for $PM_{2.5}$ data augmentation.

Model	ANEncoded	AAN
Training Epochs	100	100
Learning rate	0.00001	0.00001
Optimizer	Adam	Adam
Number of Cells	64	64, 64
Batch Size	512	512
Activation Generator	ReLU	ReLU, ReLU
Activation Discriminator	ReLU, Sigmoid	ReLU, Sigmoid

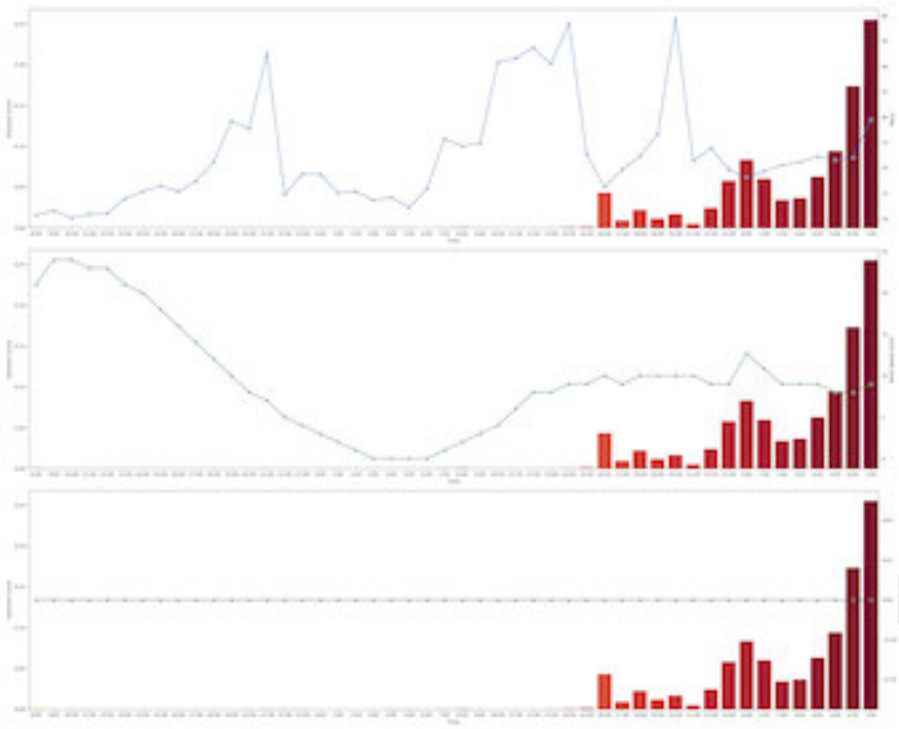

FIGURE 12. Attention scores for the different timesteps the attention block gives in the AAN generator when there is no precipitation.

TABLE 5. TSTR MSE of data augmentation models and the basic LSTM model.

Models	Basic LSTM	ANEncoded	AAN
Loss	8.92	5.91	5.55

TABLE 6. BSAM MSE when using data augmented train dataset vs original training dataset.

Models	Train MSE	Test MSE
BSAM without data augmentation	1.98	3.41
BSAM with ANEncoded	1.56	3.09
BSAM with AAN	0.95	1.27

To further validate our AAN architecture, we have conducted a two-sided Kolmogorov-Smirnov (KS) test. The KS test compares the distributions of the samples in the dataset containing real-life observations with the one of the synthetic samples. The null hypothesis stating that the two distributions are identical cannot be rejected, based on the obtained p-value of 0.815 on two large datasets of samples containing 865,000 real and synthetic data samples. Therefore, we can further claim that data augmentation does not only improve the performance of the forecasting BSAM model, but the statistical evidence confirms the AAN produces a variety of realistic new data samples i.e., high fidelity data.

We further analyze the behavior of the best-performing adversarial network AAN from the aspect of the attention mechanism employed. The distribution of attention weights over the last 24 hours of historic observations are depicted in Figures 12 and 13. Faced with a challenge of visualizing attention given to various spatiotemporal features notwithstanding the weather-related features, we have chosen to represent two scenarios regarding precipitation.

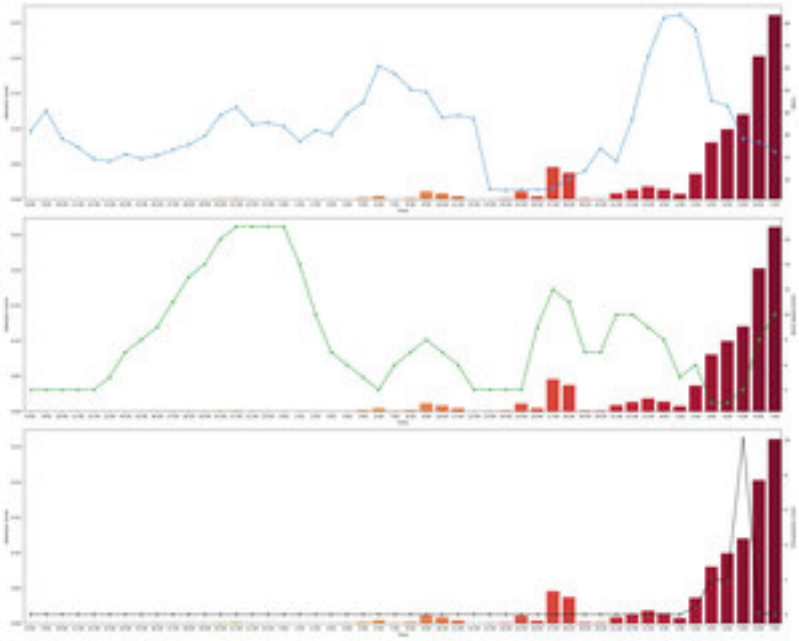

FIGURE 13. Attention scores for the different time steps the attention block gives in the AAN generator when there is precipitation.

Figure 12 consists of three plots showing the attention scores along sides: the PM2.5 observations (blue curve in the first plot), wind speed variations (green curve in the second plot); and the precipitation in the last 24-hour history window. A different 24-hour scenario for a day with precipitation is shown in Figure 13. Our insights also affirm that the BSAM model captures the multivariate context well by representing the features identified as important both, the inherent spatiotemporal and auxiliary characteristics from the last 24 time steps of history window. One can track how the attention scores varies as weather conditions are changed. The highest attention scores share this element of recency i.e. importance given to the last step in the historic observations, but the attention scores also differ in importance and relevance according to weather, wind speed and precipitation. These features are identified as important both at ..level and at .. level. It appears likely that other features beyond wind speed and precipitation affect the importance they are given when generating synthetic samples that follow the trends realistically.

Real-time monitoring and forecasting air pollution should be basis for prescribed prevention, management and alleviating the effects of air pollution that is both a global and severe problem. When facing the challenges posed by limitations of current sensor technologies i.e., missing data due to sensor malfunctions, which can significantly degrade the performances of the air pollution forecasting architectures. It is noteworthy that attention-based encoder-decoder architectures are rarely incorporated into both predictive and generative models in the domain of air pollution. There are no easy guidelines to follow when it comes to selecting the right architecture that effectively learns the model's parameters of the task at hand. Difficulties in understanding and interpreting deep learning models preclude easy replication and fair comparison between studies. The proposed architectures are general enough to be used for predictive and/or generative tasks for other pollutants or predictive and generative models conditioned on various external factors. The architectures advanced in this research are also not restricted to any particular type of external data (e.g., weather, wind conditions, location) per se as other types of external parameters might be used to condition (selective attention) the prediction or generation of a sample of a given pollutant.

VIII. CONCLUSION

The challenges air pollution forecasting faces stem from the nature of data the system is dealing with i.e. multivariate data with spatial and temporal dependencies that are often missing. Transmission disruptions and IoT sensor failures pose a serious threat to the validity of the assertions drawn from data. Addressing these longstanding concerns regarding missing data that degrade the performance of the prediction models warrants an objective of equal importance that frequently complements the task of predictive analytics in many domains. This paper reports on the predictive and generative models that are part of our general framework for IoT air pollution monitoring and forecasting system. The proposed air pollution forecasting models are based on LSTM encoder-decoder architecture. Bidirectional, stacked and attention-based LSTM encoder and decoder variants were explored and evaluated. The models can be conditioned on multiple auxiliary factors that are highly correlated with pollution; in our case study on PM2.5 prediction for the city

of Skopje, the weather condition, timeframe and location information were taken into account.

Our insights are in accordance with the related research regarding training data sizes being one of the key factors that affect the predictive performance of deep learning architectures. Data augmentation techniques are viewed as a general solution to address the issues related to the quantity and quality of training datasets. We address the challenges of missing sensor data by proposing four models for data augmentation of time series data by utilizing state-of-the-art adversarial networks. We have proposed and evaluated two novel adversarial architectures, one adding an embeddings layer to the discriminator, and the other using attention-based LSTM as encoder and decoder networks. By putting the focus selectively on the most relevant elements, when learning temporal dependencies inherent to sensor measurement sequence, the performance gains were obtained for data augmentation as well as the overall performance of the forecasting model.

As a next step, we plan to apply our approach to case studies in other cities and regions to further validate the effectiveness and versatility of our proposed architecture. While our current case study focused on particulate matter prediction for Skopje, given that we developed our framework with generality in mind, we aim to explore the performance of our models in other urban areas with different characteristics and environmental challenges. In addition, we plan to extend our approach to other time-series related problems in various domains, such as weather prediction and energy consumption forecasting, to contribute to the development of practical solutions for real-world problems.

REFERENCES

[1] J. Kalajdjieski, B. R. Stojkoska, and K. Trivodaliev, "IoT based framework for air pollution monitoring in smart cities," in *Proc. 28th Telecommun. Forum (TELFOR)*, Nov. 2020, pp. 1–4.

[2] J. Kalajdjieski, G. Mirceva, and S. Kalajdziski, "Attention models for $PM_{2.5}$ prediction," in *Proc. IEEE/ACM Int. Conf. Big Data Comput., Appl. Technol. (BDCAT)*, Dec. 2020, pp. 1–8.

[3] J. Kalajdjieski, E. Zdravevski, R. Corizzo, P. Lameski, S. Kalajdziski, I. M. Pires, N. M. Garcia, and V. Trajkovik, "Air pollution prediction with multi-modal data and deep neural networks," *Remote Sens.*, vol. 12, no. 24, p. 4142, Dec. 2020.

[4] G. Corani and M. Scanagatta, "Air pollution prediction via multi-label classification," *Environ. Model. Softw.*, vol. 80, pp. 259–264, Jun. 2016.

[5] X. Li, L. Peng, Y. Hu, J. Shao, and T. Chi, "Deep learning architecture for air quality predictions," *Environ. Sci. Pollut. Res.*, vol. 23, no. 22, pp. 22408–22417, Nov. 2016.

[6] J. Fan, Q. Li, J. Hou, X. Feng, H. Karimian, and S. Lin, "A spatiotemporal prediction framework for air pollution based on deep RNN," *ISPRS Ann. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 2, pp. 15–22, Oct. 2017.

[7] T. Li, H. Shen, Q. Yuan, X. Zhang, and L. Zhang, "Estimating ground-level $PM_{2.5}$ by fusing satellite and station observations: A geo-intelligent deep learning approach," *Geophys. Res. Lett.*, vol. 44, no. 23, pp. 1–10, 2017.

[8] X. Yi, J. Zhang, Z. Wang, T. Li, and Y. Zheng, "Deep distributed fusion network for air quality prediction," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2018, pp. 965–973.

[9] Y. Zheng, X. Yi, M. Li, R. Li, Z. Shan, E. Chang, and T. Li, "Forecasting fine-grained air quality based on big data," in *Proc. 21st ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2015, pp. 2267–2276.

[10] I. Kök, M. U. Simsek, and S. Özdemir, "A deep learning model for air quality prediction in smart cities," in *Proc. IEEE Int. Conf. Big Data*, Dec. 2017, pp. 1983–1990.

[11] Z. Qi, T. Wang, G. Song, W. Hu, X. Li, and Z. Zhang, "Deep air learning: Interpolation, prediction, and feature analysis of fine-grained air quality," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 12, pp. 2285–2297, Dec. 2018.

[12] M. Korunoski, B. R. Stojkoska, and K. Trivodaliev, "Internet of Things solution for intelligent air pollution prediction and visualization," in *Proc. 18th Int. Conf. Smart Technol.*, Jul. 2019, pp. 1–6.

[13] D. Liu, S. Lee, Y. Huang, and C. Chiu, "Air pollution forecasting based on attention-based LSTM neural network and ensemble learning," *Expert Syst.*, vol. 37, no. 3, pp. 1–12, Jun. 2020.

[14] W. Cheng, Y. Shen, Y. Zhu, and L. Huang, "A neural attention model for urban air quality inference: Learning the weights of monitoring stations," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, 2018, pp. 1–12.

[15] C. Yozgatligil, S. Aslan, C. Iyigun, and I. Batmaz, "Comparison of missing value imputation methods in time series: The case of Turkish meteorological data," *Theor. Appl. Climatol.*, vol. 112, nos. 1–2, pp. 143–167, Apr. 2013.

[16] S. Moritz, A. Sardá, T. Bartz-Beielstein, M. Zaefferer, and J. Stork, "Comparison of different methods for univariate time series imputation in R," 2015, *arXiv:1510.03924*.

[17] N. A. Zakaria and N. M. Noor, "Imputation methods for filling missing data in urban air pollution data for Malaysia," *Urbanism. Arhitectura. Constructii*, vol. 9, no. 2, p. 159, 2018.

[18] S. Rani and A. Solanki, "Data imputation in wireless sensor network using deep learning techniques," in *Data Analytics and Management*. Cham, Switzerland: Springer, 2021.

[19] T. Kim, J. Kim, W. Yang, H. Lee, and J. Choo, "Missing value imputation of time-series air-quality data via deep neural networks," *Int. J. Environ. Res. Public Health*, vol. 18, no. 22, p. 12213, Nov. 2021.

[20] J. Toutouh, "Conditional generative adversarial networks to model urban outdoor air pollution," in *Proc. Ibero-Amer. Congr. Smart Cities*. Cham, Switzerland: Springer, 2020, pp. 90–105.

[21] J. Toutouh, S. Nesmachnow, and D. G. Rossit, "Generative adversarial networks to model air pollution under uncertainty," in *Proc. 1st Int. Workshop Adv. Inf. Comput. Technol. Syst.*, May 2021, pp. 169–174.

[22] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Comput.*, vol. 1, no. 4, pp. 541–551, Dec. 1989.

[23] T. Mikolov, M. Karafiát, L. Burget, J. Cernocký, and S. Khudanpur, "Recurrent neural network based language model," in *Proc. Interspeech*, Sep. 2010, pp. 1045–1048.

[24] B. Liu, S. Yan, J. Li, G. Qu, Y. Li, J. Lang, and R. Gu, "A sequence-to-sequence air quality predictor based on the n-step recurrent prediction," *IEEE Access*, vol. 7, pp. 43331–43345, 2019.

[25] C.-J. Huang and P.-H. Kuo, "A deep CNN-LSTM model for particulate matter ($PM_{2.5}$) forecasting in smart cities," *Sensors*, vol. 18, no. 7, p. 2220, Jul. 2018.

[26] D. Qin, J. Yu, G. Zou, R. Yong, Q. Zhao, and B. Zhang, "A novel combined prediction scheme based on CNN and LSTM for urban $PM_{2.5}$ concentration," *IEEE Access*, vol. 7, pp. 20050–20059, 2019.

[27] C. Wen, S. Liu, X. Yao, L. Peng, X. Li, Y. Hu, and T. Chi, "A novel spatiotemporal convolutional long short-term neural network for air pollution prediction," *Sci. Total Environ.*, vol. 654, pp. 1091–1099, Mar. 2019.

[28] S. Haradal, H. Hayashi, and S. Uchida, "Biosignal data augmentation based on generative adversarial networks," in *Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2018, pp. 368–371.

[29] M. Alzantot, S. Chakraborty, and M. Srivastava, "SenseGen: A deep learning architecture for synthetic sensor data generation," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun. Workshops*, Mar. 2017, pp. 188–193.

[30] C. Esteban, S. L. Hyland, and G. Rätsch, "Real-valued (medical) time series generation with recurrent conditional GANs," 2017, *arXiv:1706.02633*.

[31] G. Ramponi, P. Protopapas, M. Brambilla, and R. Janssen, "T-CGAN: Conditional generative adversarial network for data augmentation in noisy time series with irregular sampling," 2018, *arXiv:1811.08295*.

[32] L. Corti and N. Oppido, "Time-conditional generative adversarial networks for augmentation of irregularly sampled time series," Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Tech. Rep., 2019.

[33] J. Yoon, D. Jarrett, and M. van der Schaar, "Time-series generative adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 5508–5518.

[34] Y. Luo, X. Cai, Y. Zhang, and J. Xu, "Multivariate time series imputation with generative adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–12.
[35] *KDD Cup*, Assoc. Comput. Mach., New York, NY, USA, 2018.
[36] J. Kalajdjieski, M. Korunoski, B. R. Stojkoska, and K. Trivodaliev, "Smart city air pollution monitoring and prediction: A case study of Skopje," in *Proc. Int. Conf. ICT Innov.* Cham, Switzerland: Springer, 2020, pp. 15–27.
[37] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," 2014, *arXiv:1406.1078*.
[38] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3104–3112.
[39] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.
[40] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
[41] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–12.

JOVAN KALAJDJIESKI received the B.Sc. and M.Sc. degrees in computer science from the Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Skopje, North Macedonia, in 2019 and 2021, respectively. In 2019, he joined the Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje. Since 2022, he has been an Applied Scientist II with Microsoft, Vancouver, BC, Canada. He has published more than ten scientific papers in international journals and conferences. His research interests include data science, machine learning, the Internet of Things, databases, data engineering, blockchain, and cybersecurity.

KIRE TRIVODALIEV received the M.Sc. and Ph.D. degrees in computer science from Ss. Cyril and Methodius University in Skopje, Skopje, North Macedonia, in 2008 and 2014, respectively. He is currently an Associate Professor with the Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje. His current research interests include data science, complex networks, the Internet of Things, machine learning, and its application in life sciences.

GEORGINA MIRCEVA received the B.Sc. and M.Sc. degrees from the Faculty of Electrical Engineering and Information Technologies, Ss. Cyril and Methodius University in Skopje, Skopje, North Macedonia, in 2007 and 2009, respectively, and the Ph.D. degree from the Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, in 2014. In 2007, she joined the Faculty of Electrical Engineering and Information Technologies, Ss. Cyril and Methodius University in Skopje. Since 2011, she has been with the Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, where she is currently an Associate Professor. Her research interests include artificial intelligence, data science, machine learning, bioinformatics, ecoinformatics, and multimedia.

SLOBODAN KALAJDZISKI received the B.Sc., M.Sc., and Ph.D. degrees from the Faculty of Electrical Engineering and Information Technologies, Ss. Cyril and Methodius University in Skopje, Skopje, North Macedonia. He is currently a Full Professor with the Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje. He has published one book and more than 80 scientific papers in international journals and conferences. His research interests include, data science, machine learning, data engineering, bioinformatics, information systems analysis and design, and databases. He was an Editor of the 10th International ICT Innovations Conference.

SONJA GIEVSKA received the B.S. degree from Ss. Cyril and Methodius University in Skopje, Skopje, North Macedonia, in 1986, the M.S. degree from the University of Zagreb, Croatia, in 1996, and the Ph.D. degree from George Washington University, Washington, DC, USA, in 2004. She is currently a Professor of computer science with the Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje. Her research interests include machine learning, artificial intelligence, and their applications.