SN Computer Science (2025) 6:637
https://doi.org/10.1007/s42979-025-04167-y

ORIGINAL RESEARCH

Enhancing $PM_{2.5}$ Air Pollution Forecasting with Novel Random Imputation Based on Hybrid RNN-Bidirectional GRU (nRI RNN-BiGRU) Model

Naushad Ahmad[1] · **Vipin Kumar**[1]

Received: 26 May 2024 / Accepted: 26 June 2025

Abstract

The issue of air pollution is critical for both the environment and global public health. It is crucial to develop accurate forecasting methods to substantially mitigate the adverse health effects of air pollution. Missing data is common in datasets where specific observations or values are not recorded. To address the problem of missing data in air quality datasets, we used a novel random imputation (nRI) method. This method accurately captures temporal dependencies of air pollution and focuses on continuously missing completely at random (MCAR) and forecasting $PM_{2.5}$ concentrations. This method accurately captures temporal dependencies of air pollution to focus on continuously MCAR and forecasting $PM_{2.5}$ concentrations. The Central Pollution Control Board provided the data in this study. Two-step methods for managing missing data follow a specific approach. In the first step, outliers are tackled by replacing them with statistically valid minimum and maximum values determined by the interquartile range (IQR). In the second step, cells that contain NaN (Not a Number) values are filled using random samples drawn from the distribution of the corresponding feature. The proposed (nRI RNN-BiGRU) model outperforms traditional deep learning models in $PM_{2.5}$ forecasting. It achieves a 27.8792 unit lower RMSE than conventional models and improves the RÂ² score by 0.506. The model also demonstrates significant error reductions across key performance metrics, with a 16.75% decrease in MAE, a 20.07% reduction in MSE, and a 10.03% improvement in MAPE compared to CNN, among others. The experimental results confirm that according to the Friedman ranking, the nRI RNN-BiGRU model consistently ranks as the most optimal model. These findings underscore its effectiveness in air pollution forecasting, supporting proactive environmental protection and public health strategies. Our findings underscore the urgency of the air pollution issue, indicating a likely increase in $PM_{2.5}$ concentration levels. The potential health risks associated with fine particulates $PM_{2.5}$, such as respiratory infections, asthma, and heart disease, further highlight the need for effective strategies for environmental protection and public health. It is, therefore, imperative to take timely, effective measures to address this issue and safeguard public health and well-being.

Keywords Forecasting · $PM_{2.5}$ · Air quality · Missing data imputation · Deep learning · RNN-BiGRU · Respiratory diseases

Introduction

Air pollution stands as a pervasive challenge in today's world, casting a shadow over the quality of our air and the health of our planet. Additionally, biodiversity, ecosystems, and ecosystem services like nitrogen cycling are negatively impacted by air pollution. Today, air pollution affects urban and rural areas both [1]. The susceptible plants, animals, people, and other creatures are now mostly at risk from the environmental contamination that results from industrial processes, which is the single aspect of the problem [2]. The concentration of air pollution is dramatically increasing daily. A single person or nation cannot resolve such global issues directly. It is answerable to everyone on the planet [3, 4]. High concentrations of air contaminants such as CO, SO_2, NO, O_3, $PM_{2.5}$, PM_{10}, and NO_2 are commonly observed [5]. A high $PM_{2.5}$ concentration in air shows that the pollutants are particularly hazardous to people [6]. The best

✉ Vipin Kumar
rt.vipink@gmail.com

Naushad Ahmad
naushad13bhu@gmail.com

[1] Computer Science and Information Technology, Mahatma Gandhi Central University, Motihari 845401, Bihar, India

Published online: 14 July 2025

illustration is particulate matter, which has a diameter of 2.5 or less and an aerodynamic diameter of no more than 10. Particulate matter is made up of both liquid droplets and solid particles that are found in the atmosphere. Elements, including elemental carbon, dust, dew, sulfides, condensates, metal particles, etc., are among the hazardous and damaging chemicals that make up $PM_{2.5}$ [7]. Long-term and short-term exposure to dangerous air pollution will cause respiratory conditions, such as breathing difficulties, cardiovascular disorders, and even lung cancer in humans. Numerous respiratory, circulatory, suffocating, irritation, and even neurological illnesses have been linked to exposure to high concentrations of particulate matter. The LRTAP is a globally recognized international agreement established within the framework of the UNECE [8]. LRTAP's principal objective is to safeguard the environment against the detrimental impacts of air pollution. This pivotal treaty was formally endorsed in 1979, signifying a significant milestone in combating air pollution on a transboundary scale [9]. The United Nations composed the five regional commissions for the UNECE. According to the WHO, air pollutants pose a significant hazard to human health, especially particulate matter, such as $PM_{2.5}$ and PM_{10}. According to estimates, $PM_{2.5}$ reduced the world's life expectancy by almost one year in 2017. According to WHO or the European Environment Agency reports, pollutant exposure raises the chance of dying young [10]. The WHO conducted various research studies revealing that air pollution causes 7 to 8 million deaths worldwide.

The Table 1 presents air pollution categories with color codes [11] and concentration ranges for PM_{10}, $PM_{2.5}$, and NO_2, along with AQI standards and health messages. Each category indicates air quality status and potential health effects on sensitive groups. The color codes range from green to Maroon, representing different levels of air pollution and corresponding health messages. The highest category, "Severe" signifies emergency conditions and health warnings for the entire population, and Table 1 highlights the national air quality index aligned with CPCB standards [12] and as utilized in recent studies (Natarajan et al., 2024) [13].

It is essential to understand the origins and quantities of these pollution contaminants. Calculating these tiny particles can be used to gauge the quality of the natural air. Real-time data on air quality is needed to manage pollution and protect humans from its effects. Missing information on ambient air pollution from databases is a common problem. Environmental epidemiology is quite concerned when data on air pollutant concentration is missing. [14]. Therefore, missing data, usually included in the gathered raw data, has become a significant obstacle for analyzing pollution patterns. The temporal and geographical mechanism of air pollution cannot be successfully captured by current research methodologies on the missing data, or they do so by concentrating on cycles with low missing rates and unpredictable missing placements [15]. Data sets on air quality frequently contain missing values since the information is gathered via sensors or by reputable organizations or government organizations. In the preliminary data processing step, recovering lost data is very difficult. As time is an inherent quantity that cannot be disregarded, this process becomes more complicated when dealing with TS data analysis [16]. The lack of automatic manipulation of missing values in various Python libraries and packages, missing values can cause a range of issues, including incorrect results and decreased accuracy, making the imputation of these values that are missing of paramount priority for improved outcomes [17].

In time series data, missingness can occur under three primary mechanisms: MCAR, MAR, and MNAR. Each mechanism poses challenges and requires suitable imputation strategies to ensure that subsequent analyses and predictions are reliable. In the context of UTS, this study compares six widely used imputation methods for handling missing data, namely simple moving average [18], mean imputation [?], and exponentially weighted moving average [19], among others. Effective handling of missing data ensures high-quality inputs for ML models in air pollution forecasting. Reliable imputation supports the development of robust models that can inform preventive strategies and enhance pollution control efforts. In recent years, regression-based models have also been applied to predict air pollution levels and assess trends [20], but their performance remains highly sensitive to data quality.

One of the most well-liked approaches, DL, uses massively scalable optimization algorithms to train a model on various data. Academic research has extensively explored

Table 1 AQI categories, standard along with their health messages of $PM_{2.5}$, PM_{10}, and NO_2 concentrations

AQI category	Colour	PM_{10}	$PM_{2.5}$	NO_2	AQI standard	Health message
Good	Green	0-50	0-30	0-40	0-50	Normal conditions
Satisfactory	Yellow	51-100	31-60	41-80	51-100	Acceptable
Moderately	Orange	101-250	61-90	81-180	101-200	Sensitive groups health effects
Poor	Red	251-350	91-120	181-280	201-300	Health affects everyone
Very Poor	Purple	351-430	121-250	281-400	301-400	Serious health effects for everyone
Severe	Maroon	431+	251+	401+	401 above	Emergency conditions

the application of statistical models in the same manner as the DL model with commonly implemented mean imputations. Several works have used DL to predict the concentration of hourly air quality. The primary goal of the cited work is to forecast O_3, $PM_{2.5}$, NO_x, and CO concentrations at a site in NCT-Delhi using the LSTM and GRU methodology, which is thought to be more effective than other DL approaches [21].

The Fig. 1 illustrates the geographical depiction of the Talkatora District Industries Center in Lucknow, as captured and presented through the utilization of ArcGIS mapping technology. Its significance as a pollution hotspot, central industrial hub, and area of regulatory interest makes it a primary target for monitoring and intervention by the CPCB [22]. The implementation of ArcGIS in this context suggests a comprehensive analysis of the area, potentially encompassing geographic and environmental data to facilitate a more well-informed comprehension of the industrial landscape in Lucknow.

This study introduces a novel random imputation (nRI) method integrated with a hybrid RNN-BiGRU model to enhance $PM_{2.5}$ air pollution forecasting. The novelty of this work lies in the strategic handling of missing data MCAR using nRI, which improves temporal pattern recognition and forecasting accuracy compared to conventional imputation methods. This research aims to develop a robust and reliable forecasting model that minimizes errors in air pollution prediction while effectively dealing with missing data challenges in large-scale air quality datasets.

Research Problem and Objective of Research

The central research problem addressed in this study is developing a robust and intelligent forecasting framework capable of capturing the temporal dependencies in

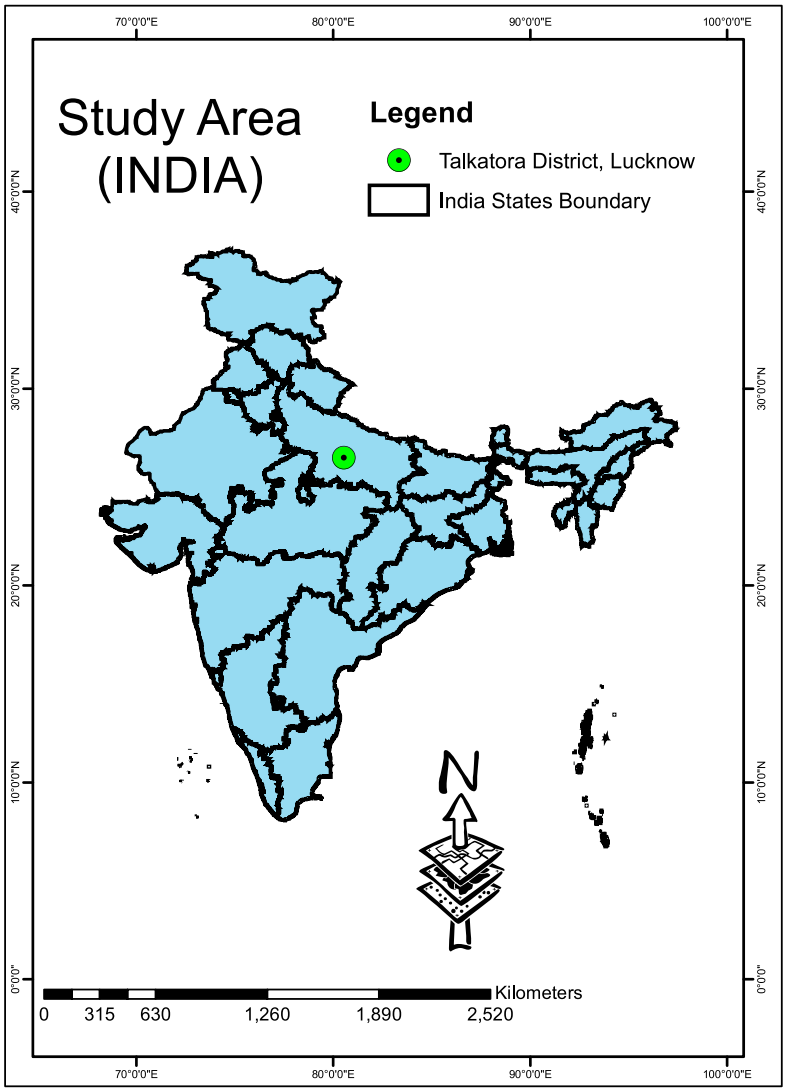

Fig. 1 The study area of the research Talkatora District Industries Center Lucknow, Uttar Pradesh

multivariate air pollution datasets while effectively handling MCAR scenarios and outliers. Existing models such as LSTM, GRU, and CNN have demonstrated promise in time-series modeling but often fall short in irregular and incomplete datasets, limiting their real-world applicability.

To overcome these limitations, this research proposes a novel random imputation method utilizing a Recurrent Neural Network with a Bi-directional GRU (nRI RNN-BiGRU) model. The objective of this work is fourfold:

- **Development of a Imputation Strategy nRI** To design and implement an efficient imputation mechanism novel random imputation that combines IQR-based outlier detection with NaN value/cell replacement using randomized feature-specific statistical sampling. This strategy aims to recover data integrity while preserving the original statistical distribution.
- **Design of the RNN-BiGRU Hybrid DL Model** To develop a hybrid deep learning architecture that integrates RNN layers with BiGRU, enabling the model to learn past and future temporal dependencies in multivariate TS data for enhanced $PM_{2.5}$ concentration forecasting.
- **Comparative Evaluation of Imputation Techniques** To investigate the impact of various traditional imputation methods on forecasting performance, including Mean, Median, KNN, BFill, and Iterative Imputation. Each technique is combined with the RNN-BiGRU model to assess its effectiveness in handling missing data. The results are compared with the proposed novel random imputation strategy, which integrates IQR-based outlier handling and NaN replacement using feature-specific random sampling.
- **$PM_{2.5}$ Pollutant Forecasting and Model Generalization** To evaluate the generalization ability of the proposed nRI RNN-BiGRU model by extending the forecasting task to other critical pollutants such as PM_{10} and NO_2. This analysis helps validate the robustness of the model across multiple pollutant types using standard performance metrics such as RMSE, MAE, MAPE, and MSE. It demonstrates its adaptability to real-world air quality prediction scenarios.

Contributions of This Research

This study proposes a novel imputation strategy that preserves the entire temporal sequence of the air pollution dataset without removing any date/time instances. Unlike conventional methods that discard incomplete entries, our approach ensures the integrity and continuity of time-series data, which is crucial for long-term forecasting accuracy. The authors collected the $PM_{2.5}$, PM_{10}, and NO_2 pollutant data directly from the CPCB. The dataset contains over **31,000 hly observations** per pollutant, with varying missing percentages (e.g., PM_{10} – 19.68%).

- **Novel Random Imputation Method:** A unique combination of IQR-based outlier handling and feature-wise randomized imputation for missing values is introduced. Outliers are replaced using the IQR's statistically valid min and max values, while NaN values are filled using random samples from the same feature distribution. This nRI technique significantly enhances model performance for noisy and incomplete air quality data.
- **RNN-BiGRU Hybrid DL Model:** The authors have developed a sophisticated hybrid RNN-BiGRU model, which effectively reveals complex patterns in data and enhances performance. This model captures the temporal dependencies of air pollutant concentrations and consistently outperforms conventional deep-learning approaches.
- **Long-Term Forecasting Capability:** The proposed nRI RNN-BiGRU model can forecast $PM_{2.5}$ levels up to one year in advance, offering a valuable early warning system. This extended horizon prediction supports urban planning and public health intervention.
- **Superior Predictive Performance:** Comparative experiments with popular deep learning models LSTM, GRU, RNN, BiLSTM, and 1DCNN demonstrate that the proposed model performs best. It attained an RMSE of **27.8792**, MAE of **15.4859**, and MAPE of **33.92** for $PM_{2.5}$ forecasting outperforming all baselines in both imputation robustness and forecast accuracy. The proposed nRI RNN-BiGRU ranked first with statistical significance p-value < 0.05.

The paper begins with the Introduction and contains the abbreviations Table 2, and the other sections are structured as follows: Sect. 2 provides a comprehensive literature review, while Sect. 3, divided into several subsections, covers the Materials and Models, including LSTM, GRU, CNN, RNN, and the proposed (nRI RNN-BiGRU) model. Sect. 4 details the Methodology, while Sect. 5 presents the Results, and finally, Sect. 6 offers the Conclusion. Declarations are mentioned at the end of the paper.

Literature Review

Younes et al. (2025) [23] evaluate two energy systems, one with an electric boiler and one with a gas boiler. While potentially more cost-effective, incorporating gas boilers still relies on fossil fuels, which can contribute to air pollution through emissions such as NOx and particulate matter. By promoting the use of renewable energy sources like

Table 2 A comprehensive list of commonly used abbreviations that enhance understanding and communication

Abbreviation	Definition	Abbreviation	Definition
PM2.5	Particulate matter 2.5 Âµm	PM_{10}	Particulate matter 10 Âµm
nRI	Novel random imputation	NO_2	Nitrogen dioxide
NaN	Not a number	BiLSTM	Bidirectional long short-term memory
LRTAP	Long-range transboundary air pollution	MAR	Missing at random
IQR	Interquartile range	UNECE	United Nations Economic Commission for Europe
GRU	Gated recurrent unit	O_3	Ozone
CNN	Convolutional neural network	WHO	World Health Organisation
RMSE	Root mean square error	AQI	Air quality index
MCAR	Missing completely at random	CO	Carbon monoxide
RF	Random Forest	MNAR	Missing not at random
ARIMA	Autoregressive Integrated Moving Average	UTS	Univariate time series
CPCB	Central Pollution Control Board	MAE	Mean absolute error
RNN	Recurrent neural network	SO_2	Sulfur dioxide
MAPE	Mean absolute percentage error	MSE	Mean squared error
EHR	Electronic health records	DI	Data imputation
ADF	Augmented dickey-fuller	KPSS	Kwiatkowski Phillips Schmidt Shin
SD	Standard deviation	PACF	Partial auto-correlation function
ACF	Auto correlation function	FF	Forward Fill
BFill	Backward Fill	KNN	K-Nearest Neighbors
ML	Machine Learning	DL	Deep Learning
TS	Time series		

photovoltaic panels, there is a push away from fossil fuel dependency, leading to a potential reduction in air pollution levels. Mohammad et al. (2023) [24] usage of fossil fuels has led to rising levels of greenhouse gases and significant air pollution, contributing to public health issues. By transitioning to renewable energy, such as photovoltaic systems, the Article posits that it can help mitigate these environmental impacts, reducing air pollutants that affect health and the environment [24]. The primary environmental sustainability Sahar et al. (2024) [25] contributing to Mashhad city air pollution is vehicular emissions, exacerbated by rapid urbanization and population growth. Expanding green spaces and investing in energy-efficient public transport can significantly reduce air pollution, improve air quality, and enhance overall urban sustainability. A recent [26] assessment of PM_{10} and $PM_{2.5}$ concentrations in four subway stations in Tehran highlights significant air quality concerns compared to outdoor levels. The average PM_{10} measurements ranged from 68 to 89 $\mu g/m^3$, while $PM_{2.5}$ levels were between 62 and 76 $\mu g/m^3$, both exceeding indoor air quality standards. Alarmingly, particulate matter levels were 1.5 to 1.7 times higher indoors. Contributing factors include high passenger traffic, continuous train operations, and inadequate ventilation. This elevated exposure increases commuter's risk of respiratory and cardiovascular diseases, emphasizing the urgent need for public health interventions.

The Table 3 provides an exhaustive overview of pollutants, their sources, and associated health effects as documented in scholarly literature. It includes $PM_{2.5}$, PM_{10}, ozone gas, methane, NO_2, and CO, along with their respective causes of pollution such as soot, dirt, liquid droplets, resuspended dust, smog, mobile combustion, landfills,

Table 3 An overview of the most prevalent air pollutants, their underlying causes, and the diseases they can trigger, accompanied by the relevant year of publication

S.No	References	Air Pollutant	Causes	Diseases	Year
1	[27]	$PM_{2.5}$	Soot, dirt, and liquid droplets	Lung carcinogen	2023
2	[28]	PM_{10}	Resuspended dust	Heart disease	2023
3	[29]	NO_2	Cars, Trucks and Buses	Cough and Wheezing	2023
4	[30]	CO	Incomplete combustion	Chest pain	2023
5	[31]	Ozone Gas (O_3)	Smog	Asthma	2024
6	[32]	Methane	Mobile Combustion, Landfills	Arrhythmia, Dizziness	2024

and incomplete combustion primarily from cars, trucks, and buses. The table also outlines diseases linked to these pollutants, encompassing lung carcinogenesis from $PM_{2.5}$, heart disease from PM_{10}, asthma triggered by ozone gas, arrhythmia, and dizziness induced by methane exposure, cough, and wheezing due to NO_2, and chest pain associated with CO inhalation. Each pollutant is tied to specific sources and consequent health ailments, summarizing the literature's findings on pollutant-related health risks.

In this literature review section, the author discusses various techniques implemented for missing data imputation, a crucial aspect of data analysis and visualization. The author has compiled a comprehensive list of literature references that are relevant to this topic and has presented them in Table 4. The table columns that are labeled as references, objective, and Imputation methods provide a clear overview of the literature that has been reviewed. The objectives of the reviewed literature are diverse and aim to achieve different goals. For example, some studies evaluate the effectiveness of various data imputation techniques, while others seek to establish a standardization for DI techniques. Additionally, some recent studies explore the most recent advancement in missing DI and related topics. Moreover, the author has written various imputation methods commonly used in practice, such as mean, median, temporal, and cross-sectional. The dataset is collected from various sources, and missing values are introduced at different rates, such as 1%, 10%, 30%, and so on. The evaluation of the model results is based on several parameters, including NRMSE, MAPE, RMSE, F1-score, and RÂ² values.

Several widely-used imputation techniques, namely: Mean (Filip Arnaut et al., 2024) [43], Median (Marziyeh Afkanpour et al., 2024)[44], K-Nearest Neighbors (Xing-Yuan Li et al., 2025) [45], Backward Fill (Zhenzhen Zhao et al., 2025)[46], Forward Fill (Xiaoying Zhang et al., 2025) [47], Interpolation (Hakjong Shin et al., 2025) [48], and Iterative (Sitao Min et al., 2025) [49] Imputation methods based on the types of missing information in the dataset. Some datasets do not contain the information in the initial time stamp or at the end, whereas almost all datasets are missing information. In the first case, BFill, FF, and interpolation imputation methods struggle because of the compulsion of the first and last information missing in Fig. 2.

It is evident in most of the literature that missing data imputation is an ongoing research area, and further studies are required to enhance our understanding and improve the existing models. The number of missing parameter NaN values significantly affects the model's effectiveness. As a result, one of the joint research questions in data science is how to create and build a good and effective data-complementing technique for decreasing data noise [50]. Many ML techniques struggle to detect TS patterns and long-term correlations in air pollution data. Only a few methods provide reliable forecasts at higher temporal resolutions, such as hourly, daily, and weekly, highlighting the need for improved forecasting solutions [5]. The link between the

Table 4 The literature survey is accompanied by citations, objectives, and methods for filling in missing data

References	Objective	Imputation Method	Year	Dataset	Missing values	Results Evaluations
[33]	Effectiveness of DI Techniques	Mean, median	2019	UCI machine learning repository	injected varying percentage	NRMSE
[34]	An innovative hybrid DI conditions	Hybrid k-NN and miceMICE	2019	marine commercial DAQ system	not specified	APE, MAPE
[35]	DI using multivariate time series	MTS generative adversarial network	2019	Toy datasets	varying percentage	Reconstruction error
[36]	An improved missing DI	Multivariate Chained Equation	2020	UCI Machine Learning Repository	not specified	RMSE
[37]	Missing DI using decision trees and fuzzy	Decision trees and fuzzy	2020	UCI Machine Learning Repository		RMSE
[38]	A Standard for DI Techniques	Mean/median	2021	69 datasets with numeric and categorical columns	1%, 10%, 30%, and 50%	F1 for classification, RMSE for regression
[39]	The most recent advances in missing DI	Temporal and cross-sectional	2022	public intensive care unit (ICU) database	Native missing 0-16 %	nRMSD
[40]	Imputing missing spatiotemporal traffic data	Dynamic Graph Convolutional	2023	two real-world traffic datasets: PeMS08 and PeMS04	10%, 30%, 50%, and 70%	MAPE, RMSE, and MAE
[41]	Missing DI using univariate techniques	Univariate imputation method	2024	water level located in Eastern Thessaloniki	1.54% to 16.66%	RMSE, MAE, MSE, R2
[42]	Missing Value Imputation in EHR	Generative adversarial network	2025	Kaggle dataset used	not specified	R2

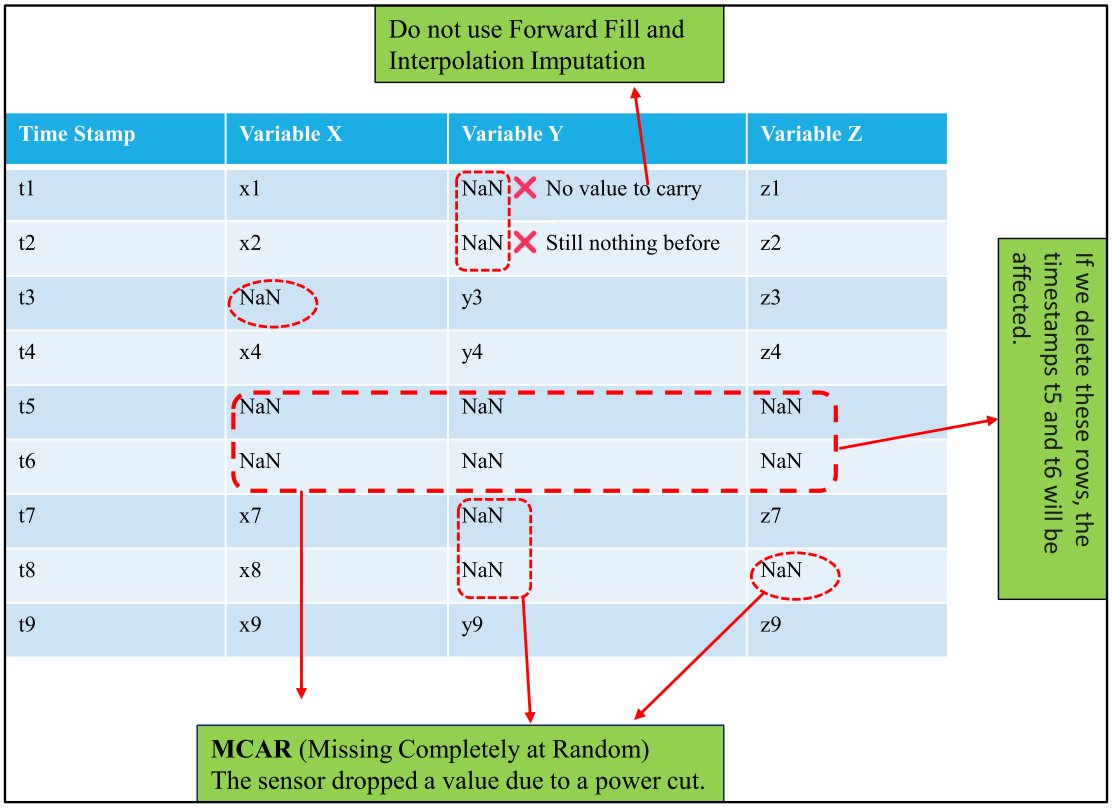

Fig. 2 Representation of missing data points in multivariate time series with missing completely at random

various meteorological factors affecting pollution and pollutant concentration. The wind direction, relative humidity, temperature, wind speed, air pressure, etc., are considered to be the most significant acknowledged characteristics with the most considerable influence on the process of pollution formation. Only if the data is correctly categorized will it be able to lessen the harm that air pollution does to human health. We are dealing with the class imbalance problem in several categorization issues. Ketu et al. utilized the learning from unbalanced data as a constant challenge [51]. The other systems require various computer environments and resources. For example, traditional machine learning may make predictions using large data platforms, but DL approaches frequently perform better on GPUs [52].

Table 5 illustrates the literature review that centers on diverse approaches, contaminants, and variables and identifies gaps in research concerning the prediction of air pollution. The review encompasses ten studies, each utilizing different techniques such as LSTM [53, 54, 62], GRU [55], regression, SVR [56], XGBoost, AdaBoost [57], RF, KNN [58], GB, MLP [59], BiLSTM [60], and ARIMA [61] to forecast levels of pollutants. Significant aspects of the prevailing methods for the performance metrics include accuracy, RMSE, MAE, MAPE, false positives, R^2 score, and SMAPE, which are most important to evaluate the model's reliability and stability. Prominent research gaps identified involve the necessity for multivariate LSTM models [53], enhanced management of extensive non-linear datasets [55], improved precision [56], and feature selection in atmospheric contexts [59]. Furthermore, certain studies stress the importance of integrating external variables [62] and enhancing the efficiency of model training, especially in centralized deep learning frameworks [60]. The table emphasizes how predictive algorithms are dynamic and how innovation is constantly needed.

Materials and Models

Lucknow, positioned at coordinates N26°50′.41 E80°53′.42, and situated at an elevation of 123 m above sea level, serves as the capital of Uttar Pradesh, a state in northern India. Geographically, it occupies a central location between the Himalayas' southern boundary and the Deccan plateau's north boundary, within the Indo-Gangetic plains notorious for their high pollution levels. Notably, Lucknow has recurrently been included in the list of the world's most polluted cities. The primary contributors to the locally prevalent $PM_{2.5}$ pollution in Lucknow are dust generated from road resuspension and construction activities. Missing data can

Table 5 The literature review focuses on models, air pollutants, and evaluation parameters, highlighting the research gap

S. No	Citations	Methods[1]	Pollutants[2]	Parameters[3]	Research gap[4]
1	[53]	LSTM	$PM_{2.5}$	Accuracy, RMSE, MAE	Multivariate LSTM for air pollution forecasting with $PM_{2.5}$ concentration
2	[54]	LSTM	multiple pollutants	R^2 score	Existing models focus on individual pollutants, perform poorly for multiple
3	[55]	GRU	$PM_{2.5}$	Accuracy, RMSE, MAE	Handling massive multivariate nonlinear datasets for air pollution prediction
4	[56]	Regression, SVR	PM	Accuracy, false positives	Need for improved accuracy
5	[57]	XGBoost, and AdaBoost	$PM_{2.5}$	MAPE, RMSE, MAE	Utilized correlation coefficient to each parameter for better accuracy
6	[58]	RF, and KNN	$PM_{2.5}$	MAE, MAPE, RMSE	Existing models show poor error rate performance in air quality forecasting
7	[59]	GB, RF, and MLP	$PM_{2.5}$, NO_2, SO_2, and CO	R^2, RMSE, and MAE	Feature selection not explored in atmosphere-related applications before
8	[60]	BiLSTM	$PM_{2.5}$ and PM_{10}	MAE, RMSE, SMAPE	Slow training speed due to centralized deep learning architecture
9	[61]	ARIMA and LSTM	$PM_{2.5}$	RMSE	Utilized deep learning model for better accuracy
10	[62]	LSTM and ANN	NO_2	MSE	Exogenous variables' impact on NO_2 forecasting performance not extensively studied

Note: Listed some of the footnotes for a clear understanding of the literature review and procedure of the same research work

[1]Taking most appropriate paper with the statistical, ML, or DL models;

[2]Listed most hazardous pollutants concerning the LRTAP;

[3]Taking most appropriate evaluation parameter according to the results;

[4]Research gap utilized to the same paper and for future work

arise under three mechanisms: MAR, MNAR, and MCAR. The easiest solution is to ignore the missing numbers, but doing so will endanger the continuity of the time series. A dataset often comprises extraneous data values, such as NaN, outliers, and noises, which can harm prediction and forecasting procedures.

The presence of these factors, NaN, outliers, and noises can lead to high error rates, reduced prediction accuracy, and increased error rates. Creating a method to identify and manage these DI techniques, including NaN, outliers, and noises, is indispensable to achieving optimal results. The data that has been gathered contains numerous values that are either missing or extreme and deviate from the other data observations with date/time. The outliers are given null values to address this issue, which signifies their position. Introducing noise into the dataset because of the missing data hurts the models' proficiency. After cleaning the noise, NaN, and outliers using nRI Fig. 3, the maximum missing of extreme feature values may impact the model's performance, like continuous missing, as these values may not conform to the expected data pattern.

The author employed two mechanisms to impute MCAR with general outliers to IQR and NaN values, which were replaced with random values drawn from the same feature distribution. The original dataset's observed values were randomly selected, and at each cycle, a missing value was substituted for it to mimic the MCAR process. The data was collected from CPCB [22] from station Talkatora District Industries Center, Lucknow - Uttar Pradesh, India. The frequency of hourly air quality data from 1 January 2020 to 24 July 2023. For the description of the data set, see the Table 6. India's Lucknow is one of the areas where air pollution has had a very negative impact. The city's bad air quality is caused by vehicle emissions, industrial pollutants, and burning crop wastes.

The Pearson correlation analysis shows in Fig. 4 the significant correlations between pollutant concentrations. The particulate matter is the strong positive correlation $r = 0.55$ between $PM_{2.5}$ and PM_{10}, which suggests a tight link between inhalable coarse particles and delicate particulate matter. On the other hand, there is less of a connection $r = 0.35$ between $PM_{2.5}$ and NO_2, suggesting a lesser relationship between nitrogen dioxide levels and delicate particulate matter. The correlation demonstrate the intricate relationships between pollutants and their surroundings, as $PM_{2.5}$ has a greater affinity for PM_{10} than NO_2. The conclusion of the correlation must be positive for better air pollution modeling.

The stationarity analysis for the pollutants $PM_{2.5}$, PM_{10}, and NO_2. The dataset described in Table 6 is used for

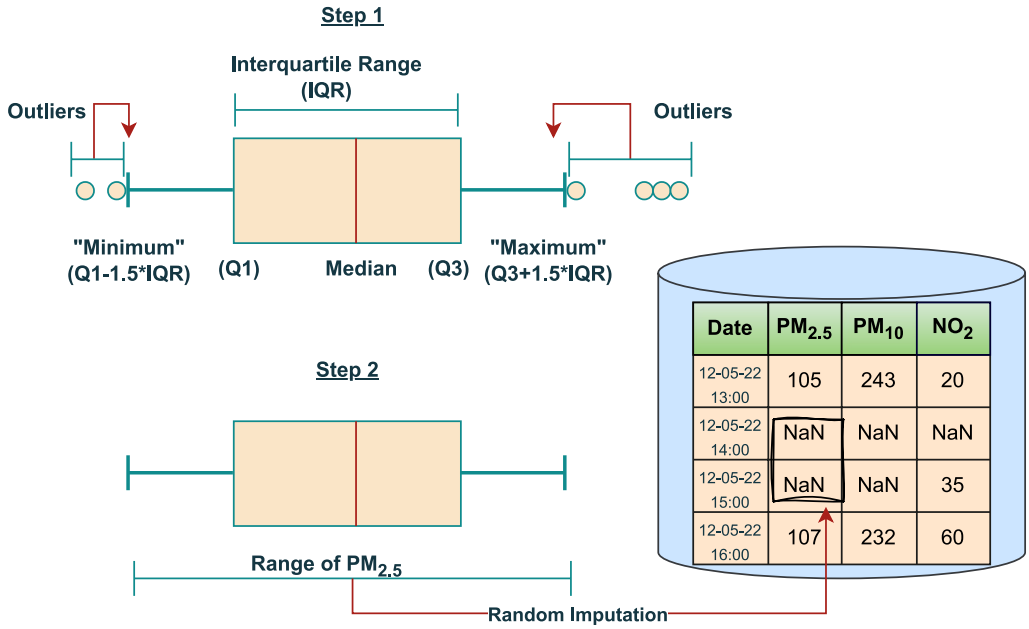

Fig. 3 Primarily steps of the Novel Random Imputation method to the implementation

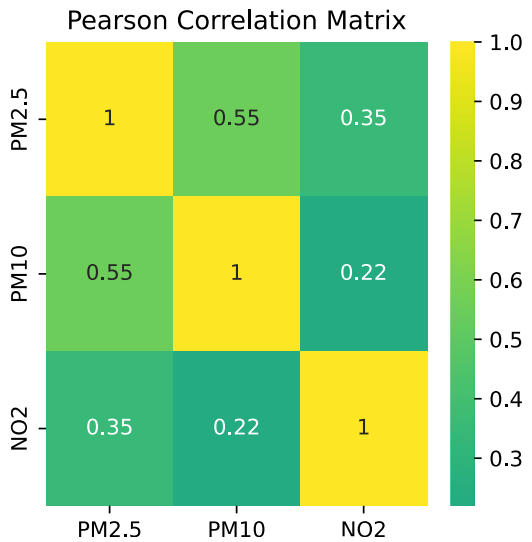

Fig. 4 The Pearson correlation of $PM_{2.5}$, PM_{10}, and NO_2 pollutant of Talkatora District Industries Center

research purposes. To calculate AQI, the author focuses on the three most common and minimum pollutants. To calculate AQI, at least three pollutants are required, with at least one being $PM_{2.5}$ or PM_{10} necessary. In this research, the researcher used $PM_{2.5}$, PM_{10}, and NO_2 pollutant concentrations. The table presents the total data points for each pollutant, along with the mean, SD, min, and max values. It also includes the first quartile, median, third quartile, and the damage beyond imputed values.

Specifically, PM_{10} exhibits 19.68% missing values, $PM_{2.5}$ has 4.63%, and NO_2 has 2.27% missing data. These levels of missingness, especially in a multivariate time-series context, can disrupt the temporal continuity [63] needed for deep learning models [64], leading to unstable gradients, degraded performance, biased predictions, or even overfitting [65] to incomplete patterns and inaccurate forecasts. Given the temporal nature of $PM_{2.5}$ data and the high correlation between pollutant levels across sequential time steps, missing entries directly impact the model's ability to learn patterns accurately. In addition, the Table 6 presents the results of the stationarity analysis using the ADF and KPSS tests to assess time series stationarity. The ADF and KPSS Statistic is compared with critical values to determine stationarity, with "Yes" indicating that the series is stationary. The results suggest that $PM_{2.5}$, PM_{10}, and NO_2 exhibit stationarity based on the ADF test, while the KPSS test indicates seasonality in these air quality parameters.

The temporal fluctuations of NO_2, PM_{10}, and $PM_{2.5}$ hourly can be effectively demonstrated by employing the statistical measures of mean and standard deviation, which can be further supplemented with a meticulous AQI color coding system Fig. 5.

The current focus in data analysis involves decomposition, which is carried out on a specific subset of the dataset $PM_{2.5}$. Specifically, the top 1000 rows are scrutinized, and a defined time frame 24 has been implemented. An additive seasonal decomposition plot has been created for 1 January 202 to 11 February 2020, with a total of 1000 rows and a period of 24. This plot displays the decomposed

Table 6 Multivariate pollution data measurement and stationarity analysis of one of the most polluted cities in India

Measurement/Stationarity	$PM_{2.5}$	PM_{10}	NO_2
Count	31224	31224	31224
Mean	91.46	186.51	40.02
Std	58.96	110.66	24.59
Variance	3473.92	12245.76	604.78
Min	0.00	0.00	0.00
25%	45.14	99.47	20.67
Median (50%)	79.37	166.38	35.02
75%	126.83	255.62	54.62
Max	255.83	467.45	109.19
Range	255.83	467.45	109.19
IQR	81.69	156.15	33.95
Skewness	0.73	0.68	0.77
Kurtosis	0.12	−0.10	−0.05
Missing	1447	6146	711
Missing%	4.63	19.68	2.27
ADF Statistic	−8.6296	−7.8283	−7.6846
ADF p-value	5.8010e−14	6.3918e−12	1.4724e−11
ADF 1%	−3.4305	−3.4305	−3.4305
ADF 5%	−2.8616	−2.8616	−2.8616
ADF 10%	−2.5668	−2.5668	−2.5668
ADF Stationary	Yes	Yes	Yes
KPSS Statistic	1.240	3.7152	4.4485
KPSS p-value	0.01	0.01	0.01
KPSS 1%	0.739	0.739	0.739
KPSS 5%	0.463	0.463	0.463
KPSS 10%	0.347	0.347	0.347
KPSS Stationary	Yes	Yes	Yes

components, including the observed, trend, seasonal, and residual components Fig. 6. Upon analysis of the data, it is evident that the trend of the data set is consistently fluctuating and increasing over time. Furthermore, the regular seasonal pattern was observed in the data set for a period of 24. The residual component, also known as statistical noise, remains after accounting for both seasonality and trends. Interestingly, no clear pattern is visible when examining the residuals graphic. The white noise observed in the data set is a stationary time series or random process with no auto-correlation, indicating that the data series lacks any predictable patterns, and each data point is independent. The data decomposition analysis provides valuable insights into the trends and seasonality of the data set, as well as the presence of statistical and white noise.

Using employing data decomposition techniques, with a particular focus on the topmost 1000 rows of $PM_{2.5}$ and utilizing a designated time frame of 24, along with scrutinizing the ACF and PACF plot lags up to 100, can yield advantageous insights about the implicit patterns and the core inherent relationships that exist within the data set. Moreover, this approach also enables the depiction of 95% confidence intervals. The auto-correlation that falls within confidence intervals is attributed to random noise Fig. 7. In time series analysis, partial auto-correlation portrays the association between the current and previous $PM_{2.5}$ values. This presumption serves as the basis for creating a linear regression model. The hypothesis of the data lacking a unit root, indicating stationarity or the absence of time-dependent structure, is rejected by a significant value of less than the defined significance level of 0.05. Consequently, rejecting

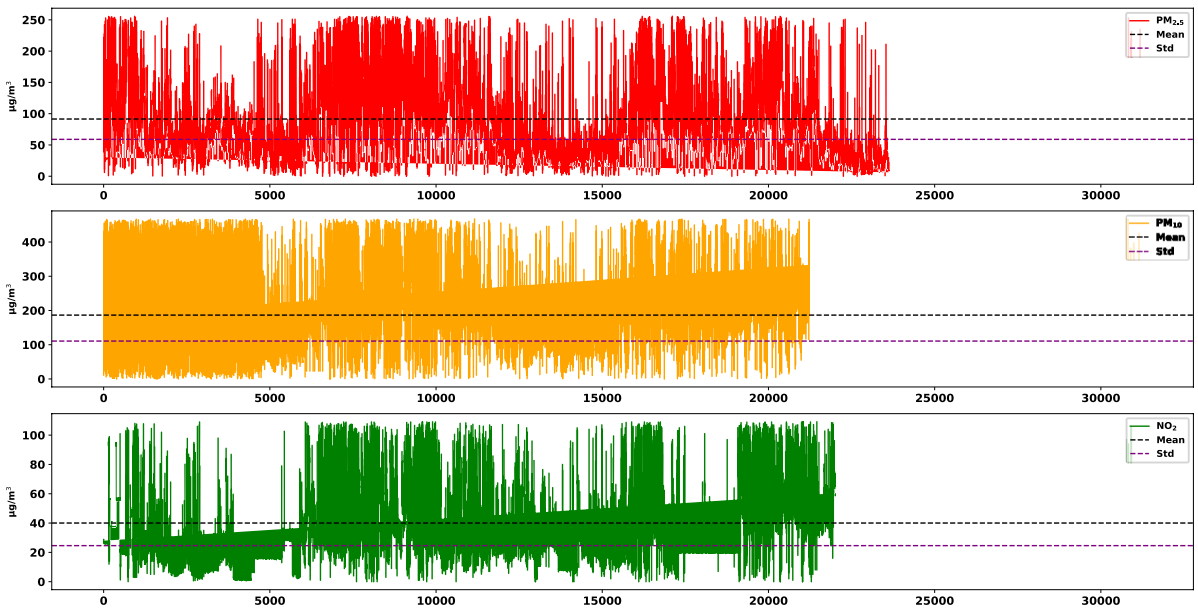

Fig. 5 Hourly fluctuations in NO_2, PM_{10} and $PM_{2.5}$ over time to mean and standard deviation with AQI color code

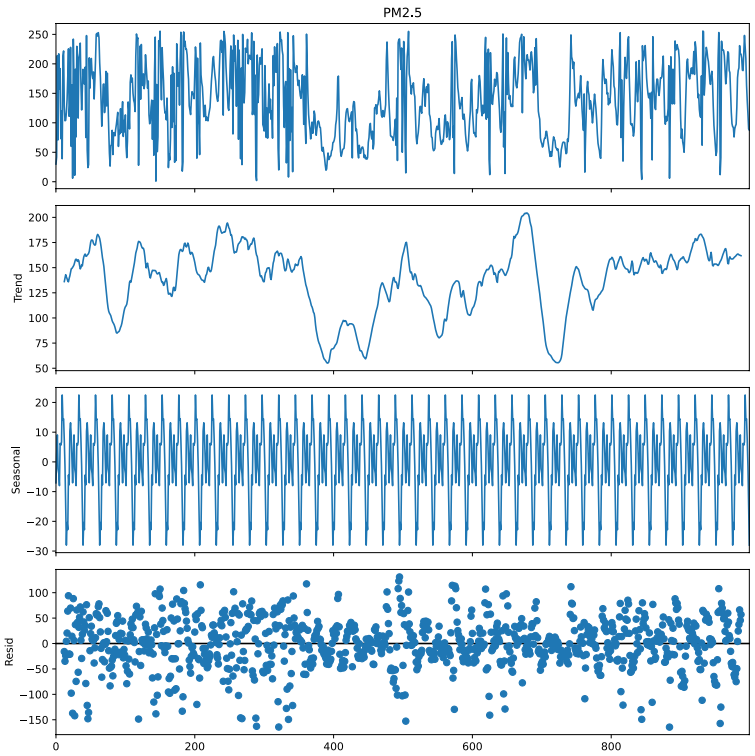

Fig. 6 Data decomposition, specifically from the top 1000 rows of $PM_{2.5}$, with a specified period of 24

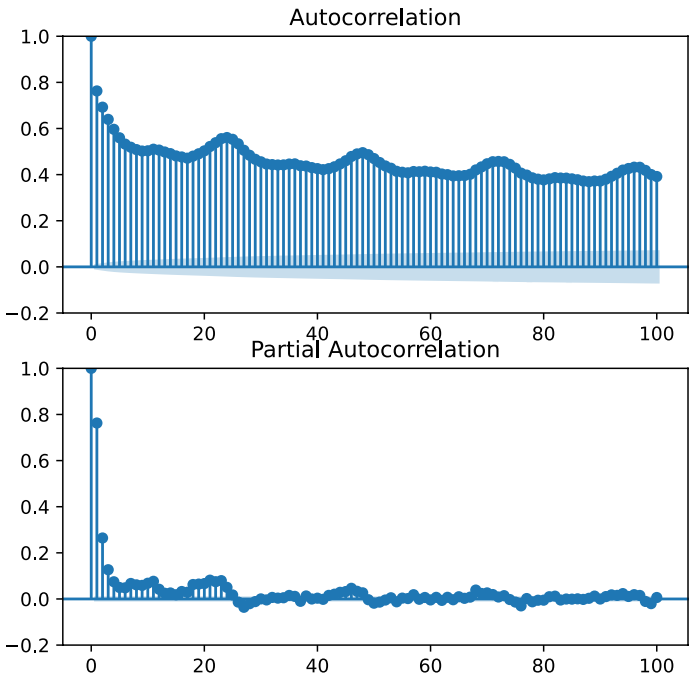

Fig. 7 Autocorrelation and partial autocorrelation of the $PM_{2.5}$ to the lags of 100

the null hypothesis (h_0) enables us to deduce that the series is stationary. The regular seasonal repetition pattern got over the lag 24 periods.

The current study presents a detailed account of the experimental procedure for the nRI in a comprehensive framework. The framework nRI starts with collecting the data set from CPCB [22], followed by preprocessing. The preprocessing phase involves the hectic task of arranging the data systematically. When the dataset comprises outliers, the author recommends replacing them using IQR without replacing NaN. To replace the NaN values, the author utilizes a specific method that generates random values within

the range of a particular column to the row. This method effectively fills the gaps within the dataset, which is crucial in ensuring the data is ready for analysis. The imputed dataset goes for the training model. The presented experimental procedure provides a valuable framework for researchers using the nRI approach in their studies Fig. 8.

Algorithm 1 The proposed Novel Random Imputation RNN-Bidirectional GRU (nRI RNN-BiGRU) model for $PM_{2.5}$ time series air pollution forecasting

Require: Feature set $X_f = \{x_{i1}, x_{i2}, \ldots, x_{im}\}_{i=1}^{n}$
Ensure: $\hat{y}_{n+k}$ (final forecasting of proposed nRI RNN-BiGRU model)
Initialization: Check if x_{ij} is an outlier: if (x_{ij} < Lower_bound) or (x_{ij} > Upper_bound)
Do not replace the value if x_{ij} is NaN (missing value).
for $i = 1$ to m **do** // Repeat the above process for each feature x_{ij} in the dataset D
  if x_{ij} < Lower_bound **then**
    $x_{ij} \Leftarrow$ Lower_bound
  else
    $x_{ij} \Leftarrow$ Upper_bound
  end if
end for
return x_{ij}
$R(i) \Leftarrow \{r_1, r_2, \ldots, r_k\}$ // Let $R(i)$ be the set of indices corresponding to the range of columns R for row i
$x_t \Leftarrow \text{RandomSelection}(R(i))$
for $i = 1$ to n **do** // This process is repeated for each row in the dataset that contains missing values
  $x_t \Leftarrow \text{RandomSelection}(R(i))$
end for
return x_{ij}
$normalized\_x_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}$ Equation 15
$h_t = \text{Models}(x_t, h_{t-1})$ // Applying models, see equation 16
The final measured (ξ_m) as $\xi_m = \xi(y, \hat{y}, \xi_m)$ see equation 18
For the final forecasting: $\hat{y}_{n+k} = \sum_{j=1}^{n} \varphi_{n,j}(k) y_{n-j+1}$ see equ 17
return $\hat{y}_{n+k}$

LSTM Model

The LSTM network is a distinguished RNN designed to effectively capture dependencies in sequential data. Its exceptional capabilities arise from an advanced internal gating mechanism composed of three critical gates: the input, forget, and output. Each of these gates plays a vital role in regulating the flow of information within the cell, ensuring that the LSTM can adeptly manage memory.

The LSTM neural network ([66]) is widely recognized as an exceedingly complex model for predicting time series, owing to its sophisticated functionalities, making it one of the most advanced models currently available in time series forecasting ([67]).

Input Gate (i_t):

$$i_t = \sigma(WC_i \cdot C_t + Wh_i \cdot h_{t-1} + b_i) \quad (1)$$

Forget Gate (f_t):

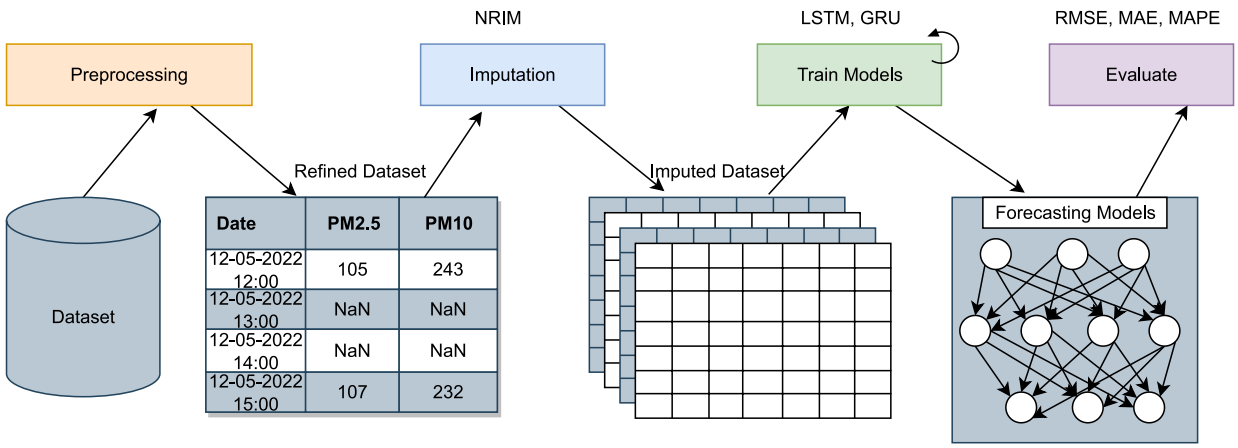

Fig. 8 The experimental procedure for the Novel Random Imputation (nRI) Method is presented in a comprehensive framework

$$f_t = \sigma(WC_f \cdot C_t + Wh_f \cdot h_{t-1} + b_r) \quad (2)$$

Output Gate (o_t):

$$o_t = \sigma(WC_o \cdot C_t + WH_o \cdot h_{t-1} + b_o) \quad (3)$$

Cell Activation (a_t):

$$a_t = f_t \cdot a_{t-1} + i_t \cdot \tanh(WC_a \cdot C_t + b_a) \quad (4)$$

Hidden State (h_t):

$$h_t = a_t \cdot \tanh(a_t) \quad (5)$$

Output (Y_t):

$$Y_t = WhY \cdot h_t + bY \quad (6)$$

Let us define the following terms: C_t represents the input, h_t is the hidden state, Y_t is the output, i_t, f_t, and o_t denote the input gate, forget gate, and output gate at time step t respectively. a_t signifies the cell activation also known as the cell state at time step t, WC_i, Wh_i, WC_f, Wh_f, WC_o, WC_a are weight matrices for input-to-gate and hidden-to-gate connections, WhY is the weight matrix for transforming the hidden state to the output, and b_i, b_r, b_o, b_a, and bY are bias vectors for the corresponding gates and output. The sigmoid function σ and the hyperbolic tangent function tanh serve as element-wise activation functions applied to the gates and cell activation, respectively Fig. 9.

The input gate controls the amount of new information added to the cell state by applying a sigmoid activation to select important features and a tanh activation to produce potential values for storage. The forget gate selectively removes outdated or irrelevant information from the cell state through a sigmoid function, ensuring only essential past information is retained. The output gate controls the information propagated to the next hidden state and final output, using a sigmoid function to regulate the output flow and a tanh function to scale the updated cell state. The flow of data through these gates and the role of each component in managing memory. A block diagram of the LSTM cell [68] provides a clear visual representation of this intricate process, illustrating the smooth flow of data through the gates and highlighting the essential functions of each component. This sophisticated mechanism empowers the LSTM to identify complex patterns in time-series data, enabling it to retain pertinent historical information while skillfully discarding irrelevant parts of the sequence. Consequently, the LSTM is an invaluable tool in various fields that require deep learning from sequential data, establishing itself as a crucial technology for addressing real-world challenges.

Gated Recurrent Unit

The GRU represents a form of RNN architecture that confronts certain constraints observed in conventional RNNs, such as the vanishing gradient issue. It is frequently applied in tasks related to NLP and other challenges involving sequential data [69]. Training the GRU model involves utilizing backpropagation through time and optimization techniques such as Adam, RMSprop, or SGD [70]. The popularity of GRU models has increased due to their capacity to capture extended dependencies and mitigate the vanishing gradient problem, rendering them advantageous for various applications [71].

The GRU is a sophisticated neural network architecture [72] that captures dependencies in sequential data while

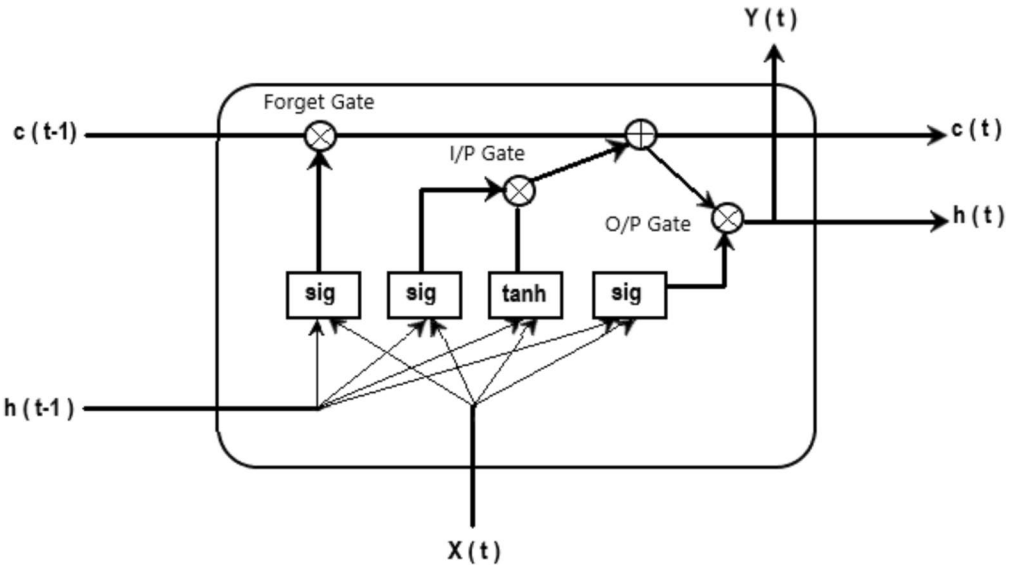

Fig. 9 Block diagram of the LSTM cell of the input gate, forget gate, and output gate

maintaining impressive computational efficiency. Unlike LSTM networks, GRUs skilfully combine the hidden and cell states into a single vector, significantly reducing the number of parameters and accelerating the training process.

The equations for updating the GRU hidden state are as follows: Calculate the update gate:

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t]) \tag{7}$$

Calculate the reset gate:

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t]) \tag{8}$$

Calculate the hidden state:

$$\tilde{h}_t = \tanh(W_h \cdot [r_t \odot h_{t-1}, x_t]) \tag{9}$$

Update the hidden state:

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \tag{10}$$

In the context provided: h_t denotes the concealed state during time step t, x_t stands for the input received at time step t, z_t refers to the gate responsible for updates, r_t represents either the reset gate or the forget gate, $\tilde{h}_t$ is the candidate concealed state, W_z, W_r, and W_h are the matrices of weights associated with the respective transformations, σ symbolizes the sigmoid function, and $\odot$ signifies the element-wise multiplication Fig. 10.

Two key components are central to the GRU's functionality: the reset and update gates. The reset gate effectively governs how much past information is discarded by modulating the influence of the previous hidden state, enabling the model to capture short-term dependencies when necessary. Conversely, the update gate determines the balance between retaining the last hidden state and incorporating new information, allowing GRUs to preserve long-term dependencies while avoiding redundant computations.

Convolutional Neural Networks (CNN)

The CNNs are deep NN designed for visual data processing [73]. CNNs use convolutional, pooling, and fully connected layers with learnable parameters. Convolutional layers apply filters to capture spatial hierarchies and local patterns in grid-like data. Pooling layers reduce dimensions and improve efficiency. Fully connected layers connect neurons to make predictions based on learned features [74]. CNNs use non-linear activation functions like ReLU to introduce complexity. 1DCNNs are specialized architectures within CNNs for one-dimensional sequential data. 1DCNNs are effective in tasks like time series analysis. They leverage convolutional layers to extract local patterns and temporal dependencies. Pooling and fully connected layers are used for downsampling and prediction [75]. They capture meaningful features in the order and temporal relationships of data points. 1DCNNs are valuable for understanding and predicting underlying patterns in sequential data.

Recurrent Neural Networks (RNN)

RNN are types of ANN tailored to process sequential data. They implement a feedback loop that allows for retaining information between time steps.In RNNs, each node maintains a hidden state that is a memory of past inputs, enabling the model to understand interdependencies and connections within sequential data [76]. This memory mechanism makes RNNs well-suited for tasks such as NLP, time series prediction, and speech recognition, where understanding context and temporal interdependencies is crucial [77]. However, traditional RNNs face challenges in capturing long-term interdependencies due to vanishing or exploding gradients. Despite these limitations, modifications such as LSTM and GRU have been developed to address these issues and

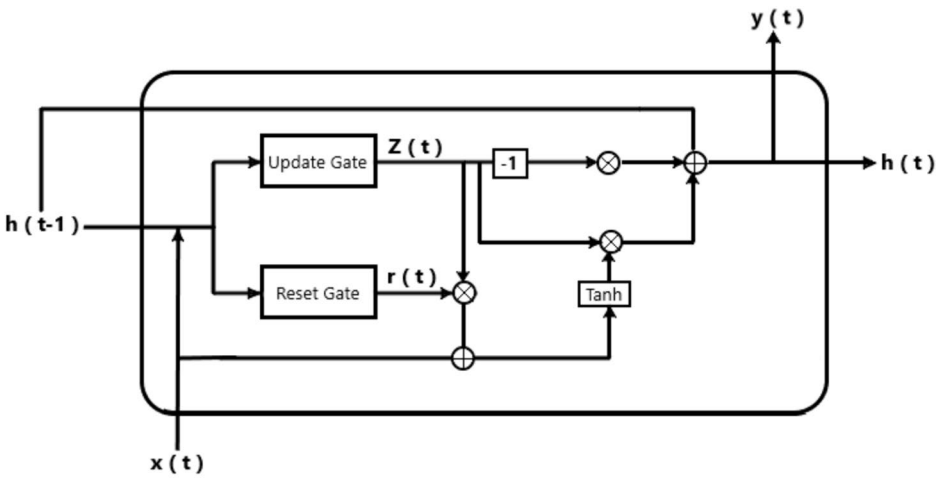

Fig. 10 Block diagram of GRU cell of the reset gate and the update gate

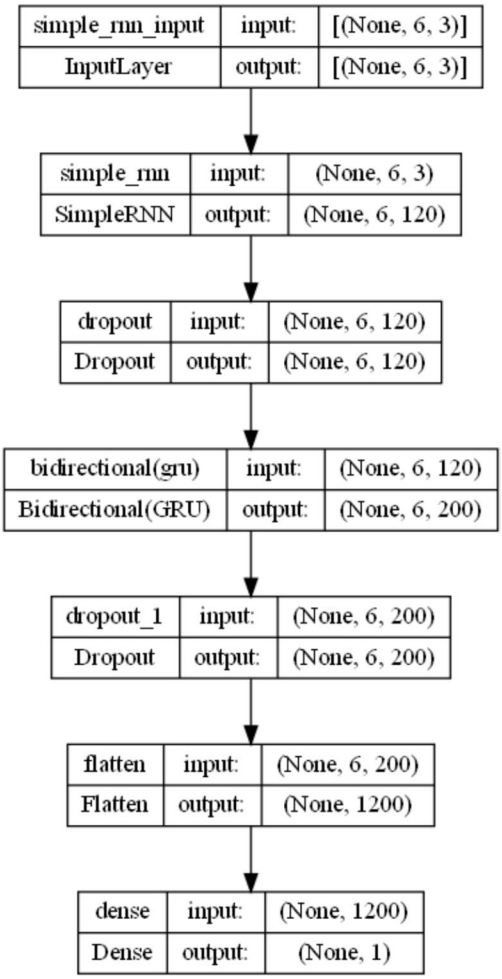

Fig. 11 Structure of proposed RNN-BiGRU model flow execution sequences

enhance the ability of RNNs to model complex temporal patterns in various domains.

Proposed (nRI RNN-BiGRU) Model

The suggested hybrid model Fig. 11 implements sequential layers, first a SimpleRNN for short temporal dependencies in the data, followed by dropout to prevent overfitting. Next, a Bidirectional GRU is added, improving both the model's forward and backward temporal dependency capture capacity. Then, again, a dropout is used for regularization. The BiGRU output is flattened to convert the 3D output to a 2D tensor. The flattened output is provided as input to a dense layer, which serves as the final layer for prediction This architecture aims to leverage the strengths of both SimpleRNN and BiGRU to model complex temporal relationships in the input data, incorporating dropout layers to enhance generalization and prevent overfitting. The model plot depicts the flow of information through these layers, highlighting the sequential arrangement of SimpleRNN, dropout, bidirectional-GRU, dropout, and finally, the flattened and dense layers.

This advanced DL model is designed for TS forecasting of air quality, specifically focusing on $PM_{2.5}$, PM_{10}, and NO_2 concentrations. With an input shape of (6, 3), it analyzes sequences of 6-time steps, each containing these three key pollutant values. The model features a SimpleRNN layer with 120 units to capture sequential dependencies and identify patterns over time, complemented by a dropout layer that reduces overfitting and enhances generalization. A bidirectional GRU layer with 200 units further strengthens the model by leveraging both past and future sequences, effectively handling fluctuations in pollution levels. Another dropout layer increases robustness by reducing reliance on specific neurons. The output from the GRU layer is flattened into a 1200-dimensional vector, which is then processed by a dense layer with a single neuron to generate the predicted concentration of a target pollutant, be it $PM_{2.5}$, PM_{10}, or NO_2. This robust architecture is well-equipped for accurate air quality forecasting, enabling reliable predictions based on historical data.

The RNN-BiGRU model, as shown in flowchart Fig. 12, is a DL approach aimed at forecasting air pollution, specifically focusing on concentrations of $PM_{2.5}$, PM_{10}, and NO_2. The process begins with data preprocessing, where outliers are handled using the Min and Max values method, and missing values are imputed using a novel random imputation method based on the IQR to ensure data consistency. After preprocessing, the dataset is split into 70:15:15 train tests and validated to optimize model learning and generalization. The RNN-BiGRU model consists of an RNN layer to capture sequential dependencies, followed by a BiGRU layer, which enhances prediction accuracy by learning from past and future contexts. Dropout layers are added to prevent overfitting, and a final dense layer produces the predicted pollutant concentration. The model undergoes hyperparameter tuning to optimize its performance, ensuring minimal validation loss. Once the proposed model is trained, it is used to predict test data and forecast air pollution levels in one year. Model performance is evaluated using standard error metrics, including RMSE, MAE, MSE, and MAPE, ensuring reliable and accurate air quality predictions. Through this structured approach, the proposed nRI RNN-BiGRU model is an effective tool for forecasting air pollution.

Methodology

The methodological framework is employed to evaluate and compare the performance of the proposed nRI RNN-BiGRU model against conventional deep learning models and traditional imputation techniques for forecasting $PM_{2.5}$

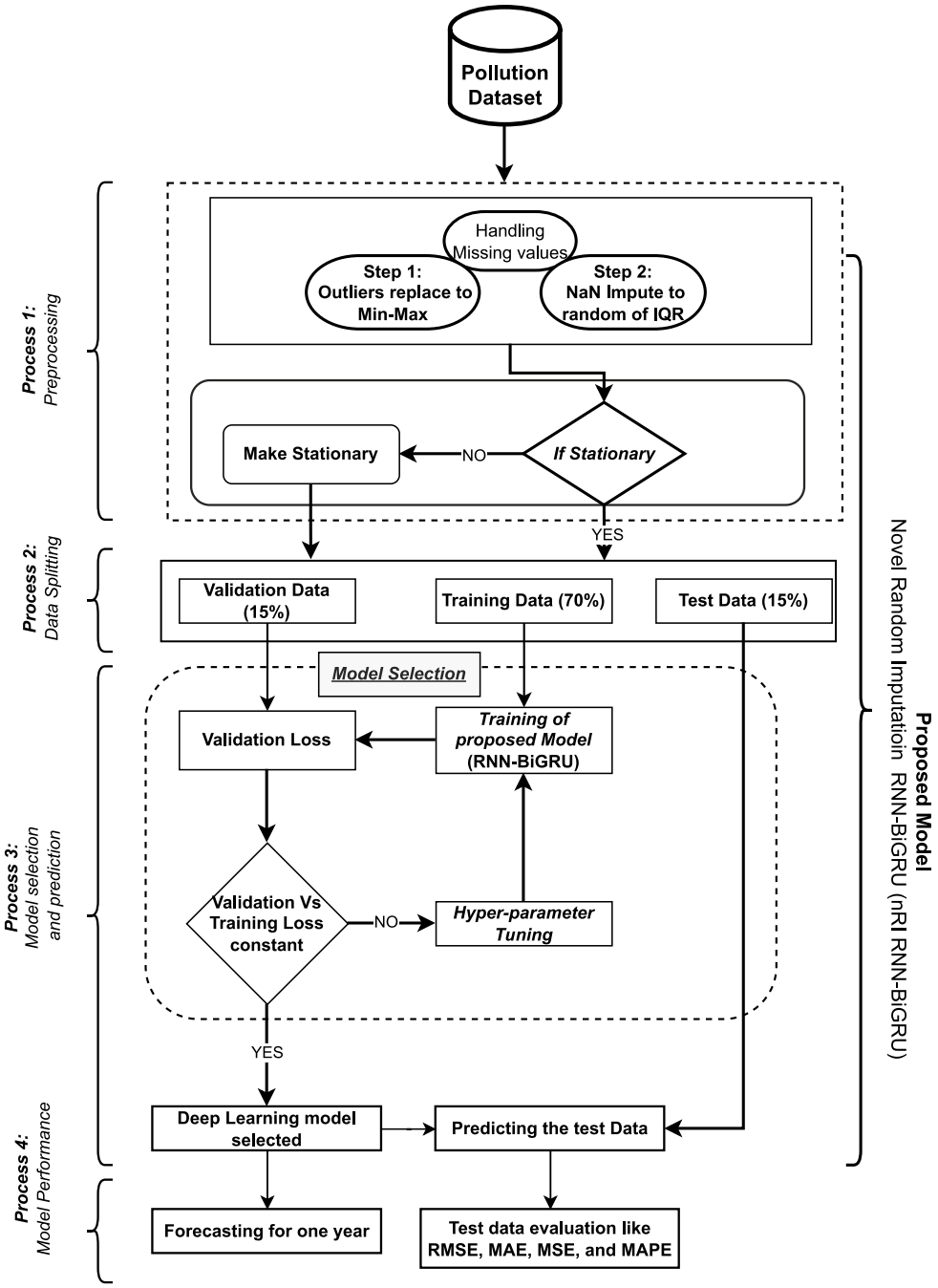

Fig. 12 The overall workflow of proposed Novel Random Imputation Recurrent Neural Networks-Bidirectional GRU (nRI RNN-BiGRU) model architecture

concentrations. The methodology encompasses data preprocessing, missing data imputation strategies, model architecture design, training procedures, and performance evaluation metrics. By systematically implementing and assessing these components, the study aims to provide a robust comparative analysis of the predictive accuracy and reliability of the proposed hybrid model in the context of air pollution time-series forecasting.

Data Collection

The collection of the data, involves obtaining a dataset D, which contains N data points $\{x_1, x_2, ..., x_N\}$ from a population. Each data point x_i is a vector with m features $\{x_{i1}, x_{i2}, ..., x_{im}\}$.

The Handling Missing Values Using Range of Columns and Random Replacement

Handling single or consecutive missing values using a range of columns and random replacement involves filling the missing values in a dataset when there are successive occurrences of missing data points along a row. This method aims to maintain the overall distribution of the data while imputing missing values with feasible estimates from the existing data. Let's a dataset D with m rows and n columns, Given a row i in the dataset D:

$$D(i) = \{x_1, x_2, ..., x_n\} \tag{11}$$

Identify the consecutive missing values in row i: Let us say we have a range of consecutive missing values starting at index j and ending at index k. Define the range of columns R. Let us say R is the set of columns that will be used to replace the missing values. Randomly select a value from the range of columns R for each missing value in the range (j to k).

$$R(i) = \{r_1, r_2, ..., r_k\} \tag{12}$$

Then, for each index t in the range (j to k), replace the missing value x_t with a randomly selected value from $R(i)$:

$$x_t = RandomSelection(R(i)) \tag{13}$$

The *RandomSelection* function randomly selects an index r_t from $R(i)$ and assigns the corresponding value to x_t:

$$RandomSelection(R(i)) = x_{r_t} \tag{14}$$

Each row with consecutive missing values goes through this process until all consecutive missing values are imputed with random values drawn from the chosen range of columns. It should be noted that this process is performed with the intention to preserve the original data distribution and variability while imputing missing values.

Normalization Using Mean and Standard Deviation

Normalization is employed to transform values of various features to a comparable range to facilitate easier learning for the model. For a given column of features x_{ij} of dataset D. Compute the mean (μ_j) and standard deviation (σ_j) for that attribute for all data points. Scale the attribute values by subtracting the mean and dividing by the standard deviation:

$$normalized\_x_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \tag{15}$$

Applying Models

In this section, we have applied all traditional deep learning models and proposed a model to determine the best forecasting accuracy. The author utilized BiLSTM, 1DCNN, GRU, LSTM, RNN, and the proposed model. When a sequence of input data x_t at time t is provided, a model processes the data and delivers an output y_t at every time step. All models can be formulated as:

$$\begin{aligned} h_t &= \text{Models}(x_t, h_{t-1}) \\ y_t &= W_h \times h_t + b_h \end{aligned} \tag{16}$$

In this context, h_t represents the hidden state at time t, while h_{t-1} is the hidden state at the previous time step $t-1$ (which is initially set to 0). Additionally, W_h signifies the weight for the output layer, and b_h denotes the corresponding bias.

The execution of the research experiment, which involved the proposed model as well as traditional deep learning models, was carried out to compare the validation loss as depicted in Fig. 13b and to showcase the training loss as Fig. 13a using the same dataset. During the experiment, the approach of early stopping was employed.

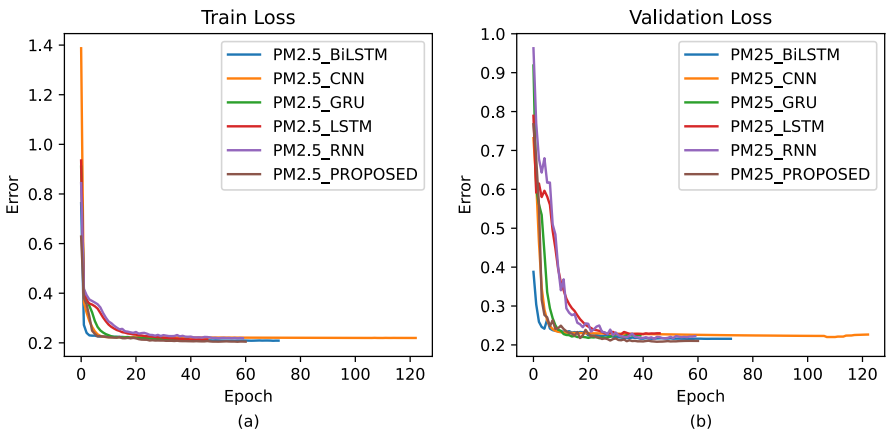

Fig. 13 Training and validation Loss comparison for all DL and proposed (nRI RNN-BiGRU) model over $PM_{2.5}$

Forecasting

Forecasting $PM_{2.5}$ air pollution concentration for the next one year using the proposed hybrid nRI RNN-BiGRU model.

$$\hat{y}_{n+k} = \sum_{j=1}^{n} \varphi_{n,j}(k) y_{n-j+1}. \tag{17}$$

where k is the value of how far into the future we make predictions, and n is the data points.

$$\xi = \xi(y, \hat{y}, \xi_m) \tag{18}$$

where ξ_m are Error measures like RMSE, MAE, MSE, and MAPE shown in the below equations. For whole algorithm Algorithm 1.

Evaluation Metrics - RMSE, MAE, MAPE, and MSE

These are performance measures for regression models. For a given set of data points with predicted values y_{pred} and true target values y_{true}, these measures are as follows:

$$RMSE = \sqrt{\frac{\sum (y_{\text{pred}} - y_{\text{true}})^2}{n}} \tag{19}$$

$$MAE = \frac{\sum |y_{\text{pred}} - y_{\text{true}}|}{n} \tag{20}$$

$$MSE = \frac{\sum (y_{\text{pred}} - y_{\text{true}})^2}{n} \tag{21}$$

$$MAPE = \frac{1}{n} \sum \left| \frac{y_{\text{true}} - y_{\text{pred}}}{y_{\text{true}}} \right| \tag{22}$$

where n is the number of holdout data set points.

Results

In this section, we perform three different types of analysis: first, quantitative analysis; second, predictive pattern and distribution analysis; and last, non-parametric statistical Friedman rank analysis.

Quantitative Analysis

In this section, we examine the exceptional quality of the traditional deep learning models and proposed models in terms of quantitative analysis. Quantitative analysis uses many methods to evaluate the model's performance. Here, the performance of regression problems in supervised ML/DL is assessed by the most common parameters like RMSE, MAE, MAPE, and MSE.

The Table 7 show that the different imputation methods like mean, median, knn, bfill, iterative, and nRI utilized the same hybrid RNN-BiGRU model. The comparison motive is to show the performance of traditional and state-of-the-art imputation methods over the nRI method. All the imputation methods are executed using the proposed model. Let us evaluate the performance of the different pollutants using various error measures. First, assess the RMSE of the $PM_{2.5}$ is proposed model 27.8792 is close to the mean and knn imputation 28.653635 and 28.95804 respectively. The RMSE value of the proposed model of PM_{10} 62.45591 is better than all the different imputation methods. The nRI RMSE value of nRI of the NO_2 13.43642 is close to the mean 11.594148 and Knn 11.720357. Similarly, all the evaluation metrics are presented with the different pollutants, various traditional imputation methods, and proposed methods.

The percentage improvement of the proposed method over traditional imputation methods is listed in the Table 8. The minimum percentage improvement of the $PM_{2.5}$ is 2.70275% to the mean imputation and a maximum of

Table 7 The performance measures of the selected pollutant of the proposed (nRI RNN-BiGRU) model with the traditional imputation methods over utilizing RMSE, MAE, MSE, and MAPE

Performance	Pollutant	Mean	Median	Knn	BFill	Iterative	Proposed nRI
RMSE	$PM_{2.5}$	28.653635	30.371923	28.95804	36.806717	41.478477	**27.8792**
	PM_{10}	78.43662	81.29937	66.44269	94.353455	115.43852	**62.45591**
	NO_2	**11.594148**	13.042486	11.720357	14.054897	16.636059	13.43642
MAE	$PM_{2.5}$	18.676916	18.92432	17.16686	23.926306	29.678127	**15.4859**
	PM_{10}	45.781338	46.039013	**32.614048**	55.25846	77.406494	33.45223
	NO_2	6.9243765	7.581353	**6.888469**	8.169521	10.365473	8.11242
MAPE	$PM_{2.5}$	34.30566	38.016052	34.678513	56.268333	78.61693	**33.92228**
	PM_{10}	37.78574	35.277332	35.64626	43.278458	54.641117	**34.61006**
	NO_2	16.316359	17.66299	**16.291214**	17.787493	24.562124	24.78002
MSE	$PM_{2.5}$	821.03076	922.45374	838.56805	1354.7345	1720.4642	**777.25**
	PM_{10}	6152.3037	6609.588	4414.6304	8902.574	13326.053	**3900.741**
	NO_2	**134.42426**	170.10645	137.36676	197.54015	276.75848	180.5375

Bold values indicate the best values in row-wise comparison

32.78635% to the iterative imputation. The minimum percentage improvement of the PM_{10} is 6.00033% to the knn imputation and a maximum of 45.89682% to the iterative imputation. The minimum percentage improvement of the NO_2 is −15.88967% to the mean imputation and a maximum of 19.23315% to the iterative imputation. Similarly, all the

Table 8 Percentage improvement of the proposed (nRI RNN-BiGRU) model over traditional imputation methods using RMSE, MAE, MAPE, and MSE

Performance	Pollutant	Mean	Median	Knn	BFill	Iterative
RMSE	$PM_{2.5}$	2.70275%	8.20733%	3.72553%	24.25513%	32.78635%
	PM_{10}	20.37404%	23.17787%	6.00033%	33.80644%	45.89682%
	NO_2	−15.88967 %	−3.02039%	−14.64173%	4.40044%	19.23315%
MAE	$PM_{2.5}$	17.08535%	18.16932%	9.79189%	35.27668%	47.82049%
	PM_{10}	26.93042%	27.33938%	−2.57%	39.46225%	56.78369%
	NO_2	−17.15741%	−7.00491%	−17.76811%	0.69895%	21.73613%
MAPE	$PM_{2.5}$	1.11754%	10.76854%	2.1807%	39.71337%	56.85118%
	PM_{10}	8.40444%	1.8915%	2.9069%	20.02936%	36.65931%
	NO_2	−51.87224%	−40.29346%	−52.10665%	−39.31148%	−0.88712%
MSE	$PM_{2.5}$	5.33241%	15.74103%	7.31223%	42.62713%	54.82324%
	PM_{10}	36.59707%	40.9836%	11.6406%	56.18412%	70.72846%
	NO_2	−34.30425%	−6.13207%	−31.42736%	8.60719%	34.76713%

Table 9 The performance measures of the selected pollutant of the proposed (nRI RNN-BiGRU) model and traditional DL model over utilizing RMSE, MAE, MSE, and MAPE

Performance	Pollutant	BiLSTM	CNN	GRU	LSTM	RNN	Propose Model
RMSE	$PM_{2.5}$	28.23212	30.54977	29.98149	30.80853	30.182	**27.8792**
	PM_{10}	63.76258	66.94911	66.1777	66.59895	68.49067	**62.45591**
	NO_2	13.37954	13.51931	13.49975	**13.30626**	13.67537	13.43642
MAE	$PM_{2.5}$	15.84844	18.08	17.55021	18.39063	17.71249	**15.4859**
	PM_{10}	34.8042	39.72214	38.86511	39.51507	41.82497	**33.45223**
	NO_2	8.155633	**7.954244**	8.186363	8.059553	8.280523	8.11242
MAPE	$PM_{2.5}$	35.35049	37.3265	35.02912	35.88219	34.67222	**33.92228**
	PM_{10}	**33.24964**	34.01362	35.64626	36.80132	35.95924	34.61006
	NO_2	25.17832	25.53941	25.35529	25.39112	24.8798	**24.78002**
MSE	$PM_{2.5}$	797.0527	933.2886	898.8895	949.1653	910.9533	**777.25**
	PM_{10}	4065.667	4482.183	4379.488	4435.42	4690.972	**3900.741**
	NO_2	179.0121	182.7717	182.2433	**177.0566**	187.0157	180.5375

Bold values indicate the best values in row-wise comparison

Table 10 Percentage improvement of error measures (RMSE, MAE, MSE, and MAPE) of all deep learning models with respect to the proposed nRI RNN-BiGRU model

Performance	Pollutant	BiLSTM	CNN	GRU	LSTM	RNN
RMSE	$PM_{2.5}$	1.265875%	9.579068%	7.540678%	10.5072%	8.259909%
	PM_{10}	2.092148%	7.194195%	5.959074%	6.633537%	9.662432%
	NO_2	−0.42335	0.616853%	0.471316%	−0.96873	1.778338%
MAE	$PM_{2.5}$	2.341065 %	16.75134%	13.3302%	18.75723%	14.37818%
	PM_{10}	4.041503%	18.74288%	16.18091%	18.12387%	25.02894%
	NO_2	0.532677 %	−1.9498	0.911479 %	−0.65168	2.072168%
MAPE	$PM_{2.5}$	4.210266%	10.03538 %	3.262877%	5.777661%	2.210757%
	PM_{10}	−3.93072	−1.72334	2.993903%	6.331268%	3.898218%
	NO_2	1.607335%	3.064501%	2.321475%	2.466095%	0.402643%
MSE	$PM_{2.5}$	2.54779%	20.07573%	15.64998%	22.1184%	17.20209%
	PM_{10}	4.228074%	14.90595 %	12.27325%	13.70713%	20.2585%
	NO_2	−0.8449	1.237522%	0.944851%	−1.92807	3.588302%

performance of the evaluation percentage matrics are presented with different pollutants compared to various traditional imputation methods. The table analysis shows that the performance of NO_2 is not much better than that of some of the imputations like mean, median, and known because the number of missing information values is much lower.

Table 9 presents a comprehensive comparison of performance measures for selected air pollutants between the proposed model and traditional deep learning models. The evaluation metrics utilized for each pollutant and the proposed model demonstrate competitive or superior performance across multiple metrics. The proposed model exhibits the lowest RMSE for $PM_{2.5}$ at 27.8792, while the BiLSTM model has the lowest RMSE among all traditional models. The RMSE values for PM_{10} and NO_2 are noted as 62.45591 and 13.43642, respectively. In terms of MAE, the proposed model achieves the lowest value of 15.4859 for $PM_{2.5}$, while the PM_{10} value is 33.45223. However, the CNN model performs better with a value of 7.954244 for NO_2. Additionally, the proposed model exhibits the lowest MAPE for NO_2 and $PM_{2.5}$ and the lowest MSE for PM_{10}. These results suggest the effectiveness of the proposed hybrid model in predicting air pollutant concentrations compared to traditional DL models, making it a promising approach for air quality forecasting.

Table 10 presents an in-depth examination of the percentage enhancement or decline in performance measurements for different deep learning models about the suggested model across various air pollutants, such as NO_2, PM_{10}, and $PM_{2.5}$. The assessment criteria include the RMSE, MAE, MAPE, and MSE. The table highlights the relative efficacy of all traditional deep learning models compared to the proposed model, where positive percentage values indicate improvement and negative values indicate decline. Notably, the proposed model consistently demonstrates superiority or competitive performance across all pollutants and metrics, thereby highlighting its effectiveness in predicting air quality compared to conventional deep learning models. The outcomes underscore the potential of the suggested model in accurately and reliably forecasting pollutant concentration.

Predictive Pattern and Distribution Analysis

Here, we have performed the predictive pattern and distributive analysis of the actual test values and predicted values of $PM_{2.5}$. The predictive pattern analysis plots multiple lines over the same period. The distributive correlation analysis works with actual and predicted values without the time innovation. The scattered plots are shown in the x-axis and y-axis terms. We can plot the actual vs. predicted values on that axis.

The current study examines the regression line using a 45-degree angle. In the overlaid scatterplots of $PM_{2.5}$ Fig. 14 is appropriately labeled to show the overall trend between actual and predicted $PM_{2.5}$. The R^2 scores of BiLSTM, CNN, GRU, LSTM, and RNN are 0.462, 0.276, 0.288, 0.25, and 0.304 respectively. The R^2 score of the proposed (nRI RNN-BiGRU) model is 0.506, which is better than all traditional DL models. The analysis indicates that the proposed model achieves the highest R^2 score of 0.506, confirming its effectiveness in predicting $PM_{2.5}$ levels.

The analysis findings indicate that the values predicted $PM_{2.5}$ by the test are highly approximate to the actual values throughout the 01-07-2023 00:00 to 24-07-2023 23:00 test prediction range. The conclusion drawn from the analysis suggests that the model performed with a high degree of accuracy, particularly in the upper regions, can be considered

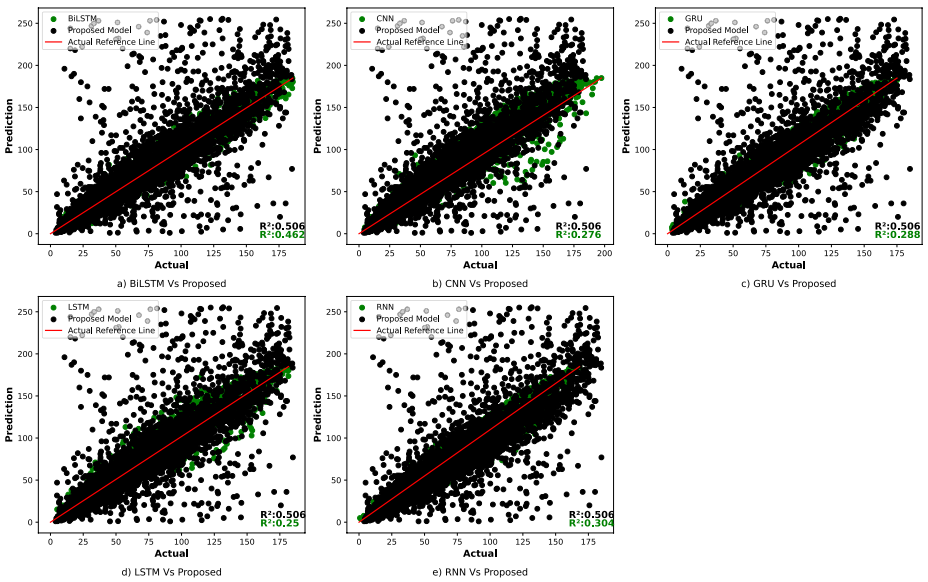

Fig. 14 Superimposed actual Vs. prediction scattered plot for proposed (nRI RNN-BiGRU) model and traditional DL models over $PM_{2.5}$ test data with R^2 score

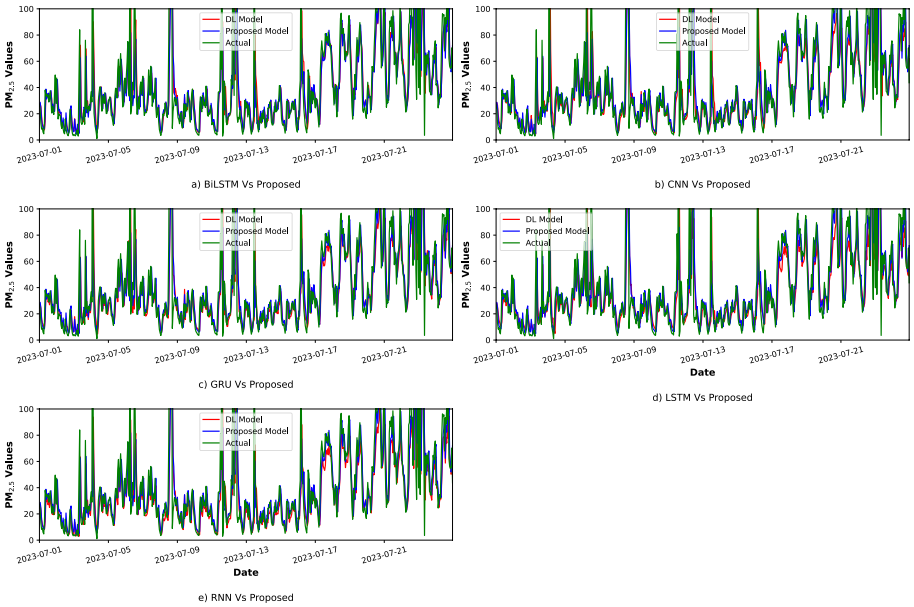

Fig. 15 Actual vs. predicted proposed (nRI RNN-BiGRU) model and traditional DL models for line plot over $PM_{2.5}$ test data

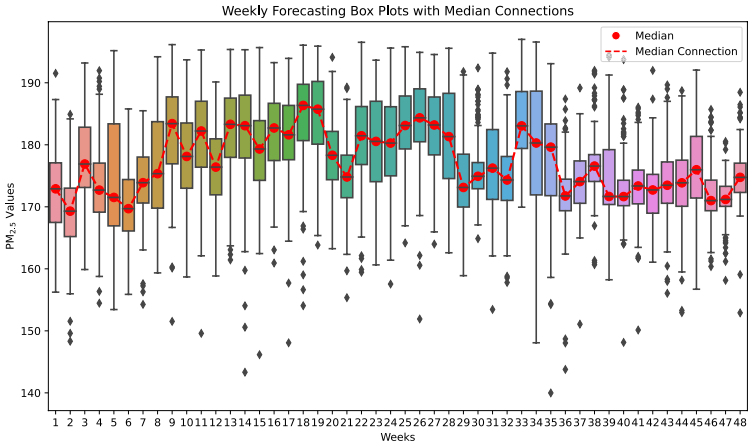

Fig. 16 One-year $PM_{2.5}$ forecasting for proposed (nRI RNN-BiGRU) model to a range of very unhealthy polluted with weakly boxplot median connecting line

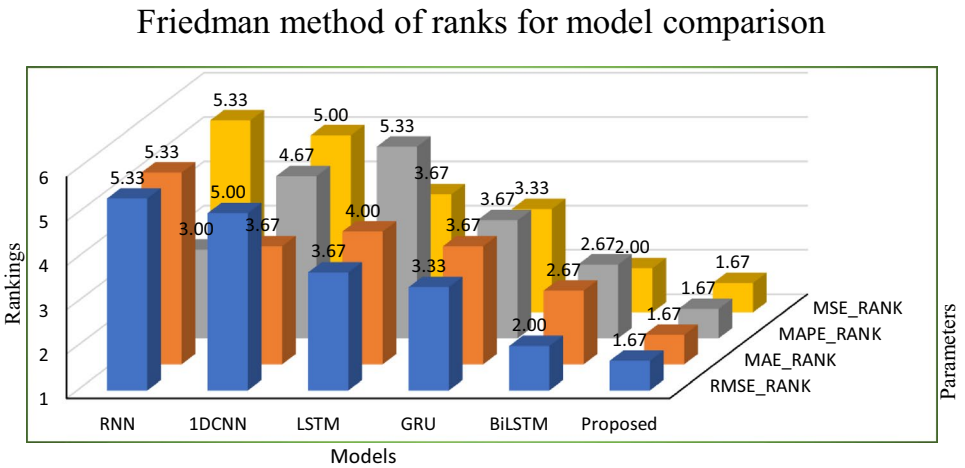

Fig. 17 The Deep Learning models and proposed (nRI RNN-BiGRU) model were evaluated using the Friedman Ranking non-parametric statistical test methodologies for all selected pollutants

satisfactory. This is indicative of the model's robustness and reliability in predicting values Fig. 15.

The Fig. 16 illustrates the one-year $PM_{2.5}$ forecasting using the proposed (nRI RNN-BiGRU) model for a range of moderately polluted conditions, presented in a boxplot format with a connecting line indicating the median. The y-axis represents the $PM_{2.5}$ concentration, while the x-axis denotes the hourly time duration 25-07-2023 00:00 to 24-06-2024 00:00 over the weakly one-year forecasting period. The boxplot visually represents the distribution of $PM_{2.5}$ concentrations, highlighting the interquartile range and the median as a central tendency measure. The connecting line facilitates tracking the median values over time. The figure offers insights into the variability and trend of $PM_{2.5}$ concentrations, aiding in assessing the model's performance in forecasting air quality under moderately polluted conditions throughout the specified timeframe. The importance of monitoring and mitigating this harmful air pollutant.

Non-Parametric Statistical Friedman Rank Analysis

The 3D bar plot in Fig. 17 exhibits the outcomes of the assessment of diverse deep learning architectures, in addition to the suggested model, utilizing the Friedman Ranking non-parametric statistical test methodologies for $PM_{2.5}$.

The evaluated architectures consist of traditional deep learning models, with their performance ranked based on four distinct measurements: RMSE, MAE, MAPE, and MSE. The lower the ranking values for these measurements, the more superior the predictive performance of the architecture. The suggested architecture, RNN-BiGRU, surpasses the other architectures in all measurements, attaining the lowest rankings in RMSE, MAE, MAPE, and MSE. The rankings for each architecture are presented for effortless comparison, with the proposed (nRI RNN-BiGRU) model architecture consistently demonstrating exceptional performance in forecasting pollutant levels.

Conclusion

Air pollution is a serious problem for the environment and public health worldwide. Forecasting accurate air pollution is very crucial. Handling missing values in massive time series data for forecasting is incredibly cumbersome. This is because removing NaN from the dataset would result in the disappearance of the date/time pattern. A proposed model called Novel Random Imputation Recurrent Neural Networks-Bidirectional Gated Recurrent Unit (nRI RNN-BiGRU) was utilized to address the issue of absent data in air quality datasets. This model effectively captures the temporal dependencies of air pollution, specifically focusing on continuous Missing Completely At Random (MCAR) data and forecasting $PM_{2.5}$. The authors used two distinct methods to address MCAR: general outliers were imputed using the Interquartile Range (IQR) method. At the same time, NaN values were replaced with random values drawn from the same feature distribution. Therefore, a stable and practical model is essential to prevent humans from experiencing the adverse health effects of $PM_{2.5}$ pollution. This study proposed the (nRI RNN-BiGRU) model for predicting and forecasting $PM_{2.5}$ concentration in Lucknow. The experimental results indicate that the proposed (nRI RNN-BiGRU) model yields the best Friedman ranking over various deep learning models. In conclusion, the nRI RNN-BiGRU method can effectively predict and forecast $PM_{2.5}$ concentration, and including PM_{10} and NO_2 data in model training can improve prediction accuracy. Non-parametric statistical tests, including **Friedman ranking** and **Holm's post hoc procedure**, were applied to rigorously assess and rank the performance of all deep learning models. The proposed **nRI RNN-BiGRU** ranked first with statistical significance p-value < 0.05, validating its general superiority across multiple pollutants and metrics. The one-year forecast reveals that the median $PM_{2.5}$ levels are expected to remain between 165 and 185 Âµg/mÂ³, categorizing air quality as "Very Poor". This suggests a serious risk to public health, with adverse effects expected for the entire population, not just sensitive groups. Fine particulates $PM_{2.5}$ may enhance the likelihood of health issues, such as respiratory infections, asthma, and heart disease, which can help devise effective strategies for environmental protection and public health.

Author Contributions Naushad Ahmad: Conceptualization of this study, Writing - Methodology, Software, Data curation, Writing - Original draft preparation, Visualization. Vipin Kumar: Validation, Formal analysis, Methodology, Investigation, Writing - Review, Editing, Supervision.

Funding No

Data availability The data is publicly available at (https://airquality.cpcb.gov.in/ccr/#/caaqm-dashboard-all/caaqm-landing).

Code availability All code was implemented in Python. Interested researchers can request access by contacting the corresponding author.

Declarations

Conflict of interest The authors declare no Conflict of interest.

Ethical approval and consent to participation Not applicable.

Consent for publication Not applicable.

References

1. Ilarri S, Trillo-Lado R, Marrodán L. Traffic and pollution modelling for air quality awareness: an experience in the city of zaragoza. SN Comput Sci. 2022;3(4):281. https://doi.org/10.1007/s42979-022-01105-0.
2. Thanigaivel S, Vinayagam S, Gnanasekaran L, Suresh R, Soto-Moscoso M, Chen W-H. Environmental fate of aquatic pollutants and their mitigation by phycoremediation for the clean and sustainable environment: a review. Environ Res. 2023. https://doi.org/10.1007/s42979-022-01105-0.
3. Biswas P, Kar N, Deb S. Ml based assessment and prediction of air pollution from satellite images during covid-19 pandemic. Multimed Tools Appl. 2024. https://doi.org/10.1007/s11042-023-18102-x.
4. Kumar V, Ahmad N. Deep learning for air quality prediction after covid-19 pandemic based on pollutant and meteorological data. Available at SSRN 2022;4292346
5. Ma J, Cheng JC, Lin C, Tan Y, Zhang J. Improving air quality prediction accuracy at larger temporal resolutions using deep learning and transfer learning techniques. Atmos Environ. 2019;214:116885. https://doi.org/10.1016/j.atmosenv.2019.116885.
6. Ji Y, Zhi X, Wu Y, Zhang Y, Yang Y, Peng T, Ji L. Regression analysis of air pollution and pediatric respiratory diseases based on interpretable machine learning. Front Earth Sci. 2023;11:1105140. https://doi.org/10.3389/feart.2023.1105140.
7. Jin X-B, Wang Z-Y, Gong W-T, Kong J-L, Bai Y-T, Su T-L, Ma H-J, Chakrabarti P. Variational bayesian network with information interpretability filtering for air quality forecasting. Mathematics. 2023;11(4):837.
8. Sarfraz Z, Sarfraz A, Sarfraz M. International partnerships and agreements for addressing air pollution. Springer; 2024.
9. Pudykiewicz J, Benoit R, Staniforth A. Preliminary results from a partial lrtap model based on an existing meteorological forecast model. Atmosphere-ocean. 1985;23(3):267–303. https://doi.org/10.1080/07055900.1985.9649229.
10. Erden C. Genetic algorithm-based hyperparameter optimization of deep learning models for pm2. 5 time-series prediction. Int J Environ Sci Technol. 2023;20(3):2959–82. https://doi.org/10.1007/s13762-023-04763-6.
11. Palanivelu S, Shree TS. Comparison of air pollutants and air quality index using spatio-temporal variation in Chennai city, Tamil Nadu. In: E3S web of conferences, Vol. 405. EDP Sciences; 2023. p. 04002. https://doi.org/10.1051/e3sconf/202340504002.
12. Aswal DK, Chandra A. Re-evaluating pollution regulation for nuclear power: addressing india's unique challenges. Environ Sci Pollut Res. 2025. https://doi.org/10.1007/s11356-025-36256-z.
13. Natarajan SK, Shanmurthy P, Arockiam D, Balusamy B, Selvarajan S. Optimized machine learning model for air quality index prediction in major cities in India. Sci Rep. 2024;14(1):6795. https://doi.org/10.1038/s41598-024-54807-1.
14. Narasimhan D, Vanitha M, et al. Machine learning approach-based big data imputation methods for outdoor air quality forecasting. J Sci Ind Res. 2023;82(03):338–47.
15. Zhang X, Zhou P. A transferred spatio-temporal deep model based on multi-lstm auto-encoder for air pollution time series missing value imputation. Future Gener Comput Syst. 2024;156:325–38.
16. Wijesekara L, Liyanage L. Mind the large gap: novel algorithm using seasonal decomposition and elastic net regression to impute large intervals of missing data in air quality data. Atmosphere. 2023;14(2):355. https://doi.org/10.3390/atmos14020355.
17. Liu D, Wang Y, Liu C, Wang K, Yuan X, Yang C. Blackout missing data recovery in industrial time series based on masked-former hierarchical imputation framework. IEEE Trans Autom Sci Eng. 2023. https://doi.org/10.1109/TASE.2023.3287895.
18. Turco M, Abatzoglou JT, Herrera S, Zhuang Y, Jerez S, Lucas DD, AghaKouchak A, Cvijanovic I. Anthropogenic climate change impacts exacerbate summer forest fires in California. Proc Natl Acad Sci. 2023;120(25):2213815120. https://doi.org/10.1073/pnas.2213815120.
19. Thakur K, Kumar H, et al. Advancing missing data imputation in time-series: a review and proposed prototype. In: 2023 international conference on emerging trends in networks and computer communications (ETNCC). IEEE; 2023. p. 53–7. https://doi.org/10.1109/ETNCC59188.2023.10284970.
20. Méndez M, Merayo MG, Núñez M. Machine learning algorithms to forecast air quality: a survey. Artif Intell Rev. 2023;56(9):10031–66. https://doi.org/10.1007/s10462-023-10424-4.
21. Barthwal A, Goel AK. Advancing air quality prediction models in urban India: a deep learning approach integrating dcnn and lstm architectures for aqi time-series classification. Model Earth Syst Environ. 2024. https://doi.org/10.1007/s40808-023-01934-9.
22. Bhawan P, Nagar EA. Central pollution control board. Central Pollut. Control Board, New Delhi, India, Tech. Rep, 2020;20–21. Available at https://cpcb.nic.in/ [Accessed: 2024-02-01]
23. Noorollahi Y, Zahedi R, Ahmadi E, Khaledi A. Low carbon solar-based sustainable energy system planning for residential buildings. Renew Sustain Energy Rev. 2025;207: 114942. https://doi.org/10.1016/j.rser.2024.114942.
24. Khah MV, Zahedi R, Eskandarpanah R, Mirzaei AM, Farahani ON, Malek I, Rezaei N. Optimal sizing of residential photovoltaic and battery system connected to the power grid based on the cost of energy and peak load. Heliyon. 2023. https://doi.org/10.1016/j.heliyon.2023.e14414.
25. Shamaee SH, Yousefi H, Zahedi R. Assessing urban development indicators for environmental sustainability. Discov Sustain. 2024;5(1):341. https://doi.org/10.1007/s43621-024-00563-1.
26. Daneshgar S, Zahedi R, Farahani O. Evaluation of the concentration of suspended particles in underground subway stations in Tehran and its comparison with ambient concentrations. Ann Environ Sci Toxicol. 2022;6(1):019–25. https://doi.org/10.17352/aest.000048.
27. Mbazima SJ. Health risk assessment of indoor and outdoor pm2. 5-bound metal (loid) s in three residential areas downwind of an active ferromanganese smelter. Air Qual, Atmos Health. 2023;16(11):2309–23. https://doi.org/10.1007/s11869-023-01409-x.
28. Ahmadian F, Rajabi S, Azhdarpoor A. Atmospheric concentrations, seasonal variations and health risk assessment of pm 2.5, pm 10, and so 2 in Tehran metropolis, Iran 2023;https://doi.org/10.21203/rs.3.rs-3441505/v1
29. Di Bernardino A, Iannarelli AM, Casadio S, Pisacane G, Siani AM. Spatial-temporal assessment of air quality in Rome (Italy) based on anemological clustering. Atmos Pollut Res. 2023;14(2):101670. https://doi.org/10.1016/j.apr.2023.101670.
30. Karountzos O, Kagkelis G, Iliopoulou C, Kepaptsoglou K. Gis-based analysis of the spatial distribution of co2 emissions and slow steaming effectiveness in coastal shipping. Air Qual, Atmos Health. 2023. https://doi.org/10.1007/s11869-023-01470-6.
31. Tang Z, Guo J, Zhou J, Yu H, Wang Y, Lian X, Ye J, He X, Han R, Li J, et al. The impact of short-term exposures to ambient no2, o3, and their combined oxidative potential on daily mortality. Environ Res. 2024;241: 117634. https://doi.org/10.1016/j.envres.2023.117634.
32. Palangi V, Macit M, Nadaroglu H, Taghizadeh A. Effects of green-synthesized cuo and zno nanoparticles on ruminal mitigation of methane emission to the enhancement of the cleaner environment.

Biomass Convers Biorefinery. 2024;14(4):5447–55. https://doi.org/10.1007/s13399-022-02775-9.
33. Jadhav A, Pramod D, Ramanathan K. Comparison of performance of data imputation methods for numeric dataset. Appl Artif Intell. 2019;33(10):913–33. https://doi.org/10.1080/08839514.2019.1637138.
34. Cheliotis M, Gkerekos C, Lazakis I, Theotokatos G. A novel data condition and performance hybrid imputation method for energy efficient operations of marine systems. Ocean Eng. 2019;188:106220. https://doi.org/10.1016/j.oceaneng.2019.106220.
35. Guo Z, Wan Y, Ye H. A data imputation method for multivariate time series based on generative adversarial network. Neurocomputing. 2019;360:185–97. https://doi.org/10.1016/j.neucom.2019.06.007.
36. Khan SI, Hoque ASML. Sice: an improved missing data imputation technique. J Big Data. 2020;7(1):1–21. https://doi.org/10.1186/s40537-020-00313-w.
37. Nikfalazar S, Yeh C-H, Bedingfield S, Khorshidi HA. Missing data imputation using decision trees and fuzzy clustering with iterative learning. Knowl Inf Syst. 2020;62:2419–37.
38. Jäger S, Allhorn A, Bießmann F. A benchmark for data imputation methods. Front Big Data. 2021;4: 693674. https://doi.org/10.3389/fdata.2021.693674.
39. Luo Y. Evaluating the state of the art in missing data imputation for clinical data. Brief Bioinform. 2022;23(1):489. https://doi.org/10.1093/bib/bbab489.
40. Kong X, Zhou W, Shen G, Zhang W, Liu N, Yang Y. Dynamic graph convolutional recurrent imputation network for spatiotemporal traffic missing data. Knowl-Based Syst. 2023;261: 110188. https://doi.org/10.1016/j.knosys.2022.110188.
41. Myllis G, Tsimpiris A, Vrana V. Short-term water demand forecasting from univariate time series of water reservoir stations. Information. 2024;15(10):605. https://doi.org/10.3390/info15100605.
42. Radosavljevic L, Smith SM, Nichols TE. A generative model for evaluating missing data methods in large epidemiological cohorts. BMC Med Res Methodol. 2025;25:34. https://doi.org/10.1186/s12874-025-02487-4.
43. Arnaut F, Đurđević V, Kolarski A, Srećković VA, Jevremović S. Improving air quality data reliability through bi-directional univariate imputation with the random forest algorithm. Sustainability. 2024;16(17):7629. https://doi.org/10.3390/su16177629.
44. Afkanpour M, Hosseinzadeh E, Tabesh H. Identify the most appropriate imputation method for handling missing values in clinical structured datasets: a systematic review. BMC Med Rese Methodol. 2024;24(1):188. https://doi.org/10.1186/s12874-024-02310-6.
45. Li X-Y, Xu Y, Zhu Q-X, He Y-L. Industrial data imputation based on multiscale spatiotemporal information embedding with asymmetrical transformer. IEEE Trans Neural Netw Learn Syst. 2025. https://doi.org/10.1109/TNNLS.2025.3527581.
46. Zhao Z, Shen G, Zhou W, Gu W, Chen C, Kong X. Convolution-aware networks for random missing traffic data imputation. Appl Intell. 2025;55(7):1–19. https://doi.org/10.1007/s10489-025-06506-1.
47. Zhang X, Zhang Y, Liu H, Li H. A bi-directional missing data imputation scheme based on cnn and short-term window sliding for thermal power units data. IEEE Access. 2025. https://doi.org/10.1109/ACCESS.2025.3554843.
48. Shin H, Park T, Jo S-K, Jung JY. Enhancing flow-through aquaculture system monitoring: a comparative study of machine learning algorithms for missing-data imputation. Aquaculture. 2025. https://doi.org/10.1016/j.aquaculture.2025.742303.
49. Min S, Asif H, Wang X, Vaidya J. Cafe: improved federated data imputation by leveraging missing data heterogeneity. IEEE Trans Knowl Data Eng. 2025. https://doi.org/10.1109/TKDE.2025.3537403.
50. Zhang Y, Wang Y, Gao M, Ma Q, Zhao J, Zhang R, Wang Q, Huang L. A predictive data feature exploration-based air quality prediction approach. IEEE Access. 2019;7:30732–43. https://doi.org/10.1109/ACCESS.2019.2897754.
51. Ketu S, Mishra PK. Scalable kernel-based svm classification algorithm on imbalance air quality data for proficient healthcare. Complex Intell Syst. 2021;7(5):2597–615. https://doi.org/10.1007/s40747-021-00435-5.
52. Chang Y-S, Abimannan S, Chiao H-T, Lin C-Y, Huang Y-P. An ensemble learning based hybrid model and framework for air pollution forecasting. Environ Sci Pollut Res. 2020;27:38155–68. https://doi.org/10.1007/s11356-020-09855-1.
53. Lavanya K, Prathik NR. Deep learning-based air pollution forecasting system using multivariate lstm. In: Artificial intelligence tools and technologies for smart farming and agriculture practices. NY: IGI Global; 2023. p. 101–14. https://doi.org/10.4018/978-1-6684-8516-3.ch006.
54. Shu Y, Ding C, Tao L, Hu C, Tie Z. Air pollution prediction based on discrete wavelets and deep learning. Sustainability. 2023;15(9):7367. https://doi.org/10.3390/su15097367.
55. Alzwy MS, Bozed KA, Maatuk AM. Air pollution prediction using model of deep learning. In: 2023 IEEE 3rd international Maghreb meeting of the conference on sciences and techniques of automatic control and computer engineering (MI-STA). NY: IEEE; 2023. p. 242–7. https://doi.org/10.1109/MI-STA57575.2023.10169675.
56. Dhanalakshmi M, Radha V. Novel regression and least square support vector machine learning technique for air pollution forecasting. Int J Eng Trends Technol. 2023;71(4):147–58. https://doi.org/10.48550/arXiv.2306.07301.
57. Kothandaraman D, Praveena N, Varadarajkumar K, Madhav Rao B, Dhabliya D, Satla S, Abera W, et al. Intelligent forecasting of air quality and pollution prediction using machine learning. Adsorpt Sci Technol. 2022. https://doi.org/10.1155/2022/5086622.
58. Kaszowski K, Godłowska J, Kaszowski W. Influence of point sources of pollution on air quality in Malopolska-first tests of a new version of forecasting of air pollution propagation system. Sci Papers Main School Fire Serv. 2023;85:59–80. https://doi.org/10.5604/01.3001.0016.3279.
59. Fang L, Jin J, Segers A, Lin HX, Pang M, Xiao C, Deng T, Liao H. Development of a regional feature selection-based machine learning system (rfsml v1. 0) for air pollution forecasting over China. Geosci Model Dev Discuss. 2022. https://doi.org/10.5194/gmd-15-7791-2022.
60. Mengara Mengara AG, Park E, Jang J, Yoo Y. Attention-based distributed deep learning model for air quality forecasting. Sustainability. 2022;14(6):3269. https://doi.org/10.3390/su14063269.
61. Suresh S, Sindhumol M, Ramadurai M, Kalvinithi D, Sangeetha M. Forecasting particulate matter emissions using time series models. Nat Environ Pollut Technol. 2023;22(1):221–8. https://doi.org/10.46488/NEPT.2023.v22i01.020.
62. González-Enrique J, Ruiz-Aguilar JJ, Moscoso-López JA, Urda D, Deka L, Turias IJ. Artificial neural networks, sequence-to-sequence Lstms, and exogenous variables as analytical tools for no2 (air pollution) forecasting: a case study in the bay of Algeciras (Spain). Sensors. 2021;21(5):1770. https://doi.org/10.46488/NEPT.2023.v22i01.020.
63. Liu S, Li X, Chen Y, Jiang Y, Cong G. Disentangling dynamics: Advanced, scalable and explainable imputation for multivariate time series. IEEE Trans Knowl Data Eng. 2025. https://doi.org/10.1109/TKDE.2025.3558405.
64. Samal R, Krishna K. Auto imputation enabled deep temporal convolutional network (tcn) model for pm2. 5 forecasting. EAI

Endorsed Trans Scalable Inf Syst. 2025. https://doi.org/10.1016/j.uclim.2021.100800.
65. Karnati H, Soma A, Alam A, Kalaavathi B. Comprehensive analysis of various imputation and forecasting models for predicting pm2. 5 pollutant in Delhi. Neural Comput Appl. 2025. https://doi.org/10.1007/s00521-025-11047-2.
66. Sherstinsky A. Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. Phys D: Nonlinear Phenom. 2020;404: 132306.
67. Shahid F, Zameer A, Muneeb M. Predictions for covid-19 with deep learning models of lstm, gru and bi-lstm. Chaos, Solitons Fractals. 2020;140: 110212.
68. Kushwah V, Agrawal P. Hybrid model for air quality prediction based on lstm with random search and bayesian optimization techniques. Earth Sci Inf. 2025;18(1):1–17. https://doi.org/10.1007/s12145-024-01514-0.
69. Huang G, Li X, Zhang B, Ren J. Pm2. 5 concentration forecasting at surface monitoring sites using gru neural network based on empirical mode decomposition. Sci Total Environ. 2021;768: 144516. https://doi.org/10.1016/j.scitotenv.2020.144516.
70. Mahjoub S, Chrifi-Alaoui L, Marhic B, Delahoche L. Predicting energy consumption using lstm, multi-layer gru and drop-gru neural networks. Sensors. 2022;22(11):4062. https://doi.org/10.3390/s22114062.
71. Li X, Ma X, Xiao F, Xiao C, Wang F, Zhang S. Time-series production forecasting method based on the integration of bidirectional gated recurrent unit (bi-gru) network and sparrow search algorithm (ssa). J Pet Sci Eng. 2022;208: 109309. https://doi.org/10.3390/s22114062.
72. Govande A, Attada R, Shukla KK. Predicting pm2. 5 levels over Indian metropolitan cities using recurrent neural networks. Earth Sci Inf. 2025;18(1):1–16. https://doi.org/10.1007/s12145-024-01491-4.
73. Selmy HA, Mohamed HK, Medhat W. A predictive analytics framework for sensor data using time series and deep learning techniques. Neural Comput Appl. 2024. https://doi.org/10.1007/s00521-023-09398-9. (**Published online**).
74. Wardana INK, Gardner JW, Fahmy SA. Estimation of missing air pollutant data using a spatiotemporal convolutional autoencoder. Neural Comput Appl. 2022;34(18):16129–54. https://doi.org/10.1007/s00521-022-07224-2.
75. Cheng X, Zhang W, Wenzel A, Chen J. Stacked resnet-lstm and coral model for multi-site air quality prediction. Neural Comput Appl. 2022;34(16):13849–66. https://doi.org/10.1007/s00521-022-07175-8.
76. Pilcevic D, Djuric Jovicic M, Antonijevic M, Bacanin N, Jovanovic L, Zivkovic M, Dragovic M, Bisevac P. Performance evaluation of metaheuristics-tuned recurrent neural networks for electroencephalography anomaly detection. Front Physiol. 2023;14:1267011. https://doi.org/10.3389/fphys.2023.1267011.
77. Li L, Yang R, Lv M, Wu A, Zhao Z. From behavior to natural language: generative approach for unmanned aerial vehicle intent recognition. IEEE Trans Artif Intell. 2024. https://doi.org/10.1109/TAI.2024.3376510.