Journal of Scientific & Industrial Research
Vol. 82, March 2023, pp. 338-347
DOI: 10.56042/jsir.v82i03.71764

Machine Learning Approach-based Big Data Imputation Methods for Outdoor Air Quality forecasting

Narasimhan D[1] & Vanitha M[2]*

[1]Department of Mathematics, [2]Department of Computer Science and Engineering, Srinivasa Ramanujan Centre, SASTRA Deemed to be University, Kumbakonam 612 001, Tamil Nadu, India

Received 11 July 2022; revised 15 September 2022; accepted 07 October 2022

Missing data from ambient air databases is a typical issue, but it is much worse in small towns or cities. Missing data is a significant concern for environmental epidemiology. These settings have high pollution exposure levels worldwide, and dataset gaps obstruct health investigations that could later affect local and international policies. When a substantial number of observations contain missing values, the standard errors increase due to the smaller sample size, which may significantly affect the final result. Generally, the performance of various missing value imputation algorithms is proportional to the size of the database and the percentage of missing values within it. This paper proposes and demonstrates an ensemble – imputation – classification framework approach to rebuild air quality information using a dataset from Beijing, China, to forecast air quality. Various single and multiple imputation procedures are utilized to fill the missing records. Then ensemble of diverse classifiers is used on the imputed data to find the air pollution level. The recommended model aims to reduce the error rate and improve accuracy. Extensive testing of datasets with actual missing values has revealed that the suggested methodology significantly enhances the air quality forecasting model's accuracy with multiple imputation and ensemble techniques when compared to other conventional single imputation techniques.

Keywords: Air quality, Big data analytics, Classification, Ensemble, Multiple imputation

Introduction

Advancements in information and communication technology have led to a dramatic increase in environmental datasets, especially in pollution control. This rise has resulted in high data analytics, and as a result, good potential learnings, rules, approaches, and patterns have been drawn from such information. Environmental contaminant monitoring is essential to exposure science development and public care practice. Government entities frequently use environmental monitors for record keeping or research purposes. Environmental health researchers use monitors to assess contaminant concentrations in the environment and link those concentrations to possible hazards and health effects.[1] Human health may be adversely affected by air pollution. Because of their links to respiratory and cardiovascular diseases air contaminants such as particulate matter and ozone have increased mortality and hospital admissions.[2,3] The air quality monitoring network provides a substantial facility for evaluating ambient air concentration conditions and establishing pollution prevention and control plans.

Despite quality assurance and quality guidelines, hour-based air concentration data from the monitoring stations are frequently offered with missing values. Missing values pose a significant challenge to information services such as online ambient air quality publishing, ensemble forecasting, and epidemiological studies.

Data can be lost in large sections due to failure, quantification periods, or a transitional power outage. Missing data is classified into 3 types: MCAR - Missing entirely At Random, MAR - Missing At Random, and NMAR - Not Missing At Random. MCAR denotes that the incomplete data mechanism is autonomous of the values of any items in the dataset, whether missing or observed. In the case of MAR, the cause of incomplete information is unconnected to missing values but may be associated with reporting the value of another variable. According to the incomplete data mechanism of air pollutant concentrations is MAR.[4] Improper datasets may introduce various problems like Loss of exactness due to a lesser amount of data, Computational problems

*Author for Correspondence
E-mail: vanitha_mu@src.sastra.edu

due to holes in the dataset, and Bias due to distortion of the data distribution.

Even though some algorithms can directly handle missing data values, such as C4.5 and KNN, the classification performance is greatly reduced when the data set contains a large amount of missing information.[5] Many approaches have been suggested to address the issue of incomplete information classification.[6,7] One strategy tries to remove missing values from the dataset and produce the results. The final result yielded by adopting this strategy may not be efficient since some valuable information is lost. Another scenario may try to fill the value with average, median, mostly repeated data, or some constant data. This strategy may provide good results compared to the previous and most adopted ones. An ML algorithm can also be used to impute the missing data values. This paper uses various imputation techniques to solve the problems related to missing data, and a detailed comparison is given in the result section.

Several incomplete data imputation methods were developed and demonstrated their unique superiorities in various scenarios over the last several decades, the majority of which are built on statistical information and machine learning concepts.[8] These approaches frequently share the concept of imputing missing values in incomplete tuples via their complete neighbours in the data set. This paper uses various imputation techniques to solve the problems related to missing data, and a detailed comparison is given in the result section.

The proposed work advocates the amalgamation of benefits of various incomplete value imputation methods, which have all been shown to have distinct superiorities in different scenarios. More specifically, it chooses some traditional incomplete data imputation strategies to fill up the existing incomplete dataset, yielding multiple complete datasets. The proposed work first fills up the partial dataset using several imputation policies to produce a cluster of filled datasets and then performs ensemble classification on those complete datasets.

Related work

There are 2 kinds of styles in the literature for dealing with incomplete data value problems: remove the case and impute with possible values. The first is the optimum method for dealing with incomplete data values, as it can remove the incomplete information record in the data set. This only applies to a restricted set of data records, resulting in a bias in classification problems. Another method is imputation, which replaces missing data attributes with feasible data attributes.[7] When the proportion of incomplete information is low, imputation is most valuable. If the percentage of missing information is too large, the outcomes lack significant variability, which could not lead to an effective framework.

Before deciding on an approach, data scientists must first determine why the information is missing.[9] Typically, removing MCAR data is safe because the outcomes will be unbiased. Although the test will be less powerful, the findings will be reliable. The missing data in MAR can be predicted using the complete observed values. Since the missing data is unknown, like MAR, incomplete values cannot be decided by the observed data in MNAR. To establish an unbiased estimate, data scientists must prototype the missing data. Clearing observations with incomplete information may result in a biased model.

Deletion strategies are the most basic and conventional procedures for dealing with incomplete data and are widely used in statistical software. List wise deletion and pair wise deletion are the standard deletion methods. In list wise deletion, also called complete case analysis, all cases with incomplete values on one or more data points are removed from the dataset. This method has the advantage of completing the remaining dataset. However, due to the absence of incomplete cases, this complete data file has a smaller sample size and power. Furthermore, there is a possibility of a biased dataset if the data is not MCAR. In most cases, the drawbacks of list wise deletion far exceed the benefits. Nonetheless, this method is still widely used in many fields of research. Furthermore, in several statistical software packages, this method is the default choice in many statistical procedures.[10]

In pair wise deletion, known as available case analysis, missing instances are deleted one by one. As a result, each case can sometimes contribute to a special assessment but not others. The sample size will be the same for some evaluations and reduced for others in this approach. The main issue with this strategy is the assumption that the MCAR mechanism produces unbiased estimates, but varying sample sizes can also cause problems in evaluating standard errors.

There are two techniques for imputing missing data: single and multiple imputations.[11] Methods for Single imputation use a precise value especially mean

or median to replace each missing information. The complete data set can be applied directly to interpret the findings on related research grounds. Multiple imputation approaches create multiple simulated data for each incomplete one to mirror the variance with the missing data. In general, a multiple imputation mechanism requires a complete assertion of the distributional form of the variable to obtain the conditional distribution of the incomplete information given the observed data.

When handling missing data: mean, median, mode, constant, and most frequent are the most commonly used methods for imputing values. When there are only a few incomplete observations, data engineers can compute the mean or average of the remaining observations. However, mean or average results can lead to a loss of variation when there are so many missing variables. This method does not rely on time series characteristics or variable relationships. This method is still principal for underestimation of the total sample variance and there is a possibility of Type I error. K-Nearest Neighbour is another simple method for filling missing samples.[12] This procedure uses the K adjacent known neighbour's value to impute missing points. Nevertheless, this technique is best suited to fill short-length gaps in missing data.

Regression techniques were used to guess data points with missing values based on the available data. Although only one variable must be recognized, this is an operative method for imputing short periods of incomplete data samples. The Expectation-Maximization (EM) algorithm has been utilized for incomplete weather data records.[13] The EM algorithms use iterative computation such as prediction and estimation to acquire maximum likelihood estimates. However, if the missing portion is not part of a recurring pattern, EM may not function as expected in the absence of a repeated pattern in a sample.

A spatial imputation scheme that feeds the missing information using data from outdoor air stations is accomplished by a low-rank matrix completion algorithm.[11] It takes advantage of high spatial association and steadiness of air pollutants spatial matrix. It divides the spatial matrix into a low-rank matrix representing the spatial association and a sparse matrix that handles the possible outliers due to measurement errors, which intensely fills the missing observations in air contaminants data sets. Thus, the pollutants space matrix must be a low-rank matrix.

Another approach used in air quality studies that record the time series characteristics of the natural monitoring data is univariate time-series imputation.[14] Last Observation Carried Forward (LOCF) is a common method for connecting data by filling in missing data gaps with the most recently observed value. The mean hourly method is another technique for imputing absent concentration levels for a fixed air monitoring station. This method employs hourly concentration levels observed at the same monitor over long periods, frequently months or a year. When the same monitor is missing data, observed hourly mean values gathered at the same monitoring station are used to impute hours.

Multivariate Imputation by Chained Equations (MICE) is a cutting-edge multiple imputation technique. To forecast missing values, MICE employs some regression models.[6] First, each incomplete field is replaced at random with one complete value from the same attribute. Each incomplete attribute is then approximated on other characteristics to construct a beer approximation for the attribute. The procedure is repeated for all insufficient features to yield a single imputed dataset. The entire procedure is repeated for N times ($N > 1$) to generate N imputed data records. Finally, the average of the N datasets is considered to produce the final imputed data samples.

In this paper, the suggested framework encourages the combined rewards of different imputation methods for the treatment of missing values, which have been recognized that they all have distinguishing advantages in different scenes. A single imputation method cannot achieve high performance at all stages and the performance relies entirely on the nature of the dataset and diversity of samples.

Multiple Imputation generally retains all the foremost advantages of single imputation and fixes its foremost disadvantages. It is highly efficient for small sample size data. From the above studies, it is also indicated that the classification accuracy gets improved when the missing values are get imputed before applying classification. To develop good forecasting with less variance and bias, the missing values should be imputed by considering sufficient diversity between the available data samples.

Experimental Work

Even though myriad approaches are there to forecast the quality level of ambient air pollution, performance enhancements are still required for today's world. The

majority of the available works do not focus on the type of information available for different contaminants and their natural influence on the ultimate result. This paper recommends a new approach for incomplete data classification that encapsulates an imputation framework to improve the outcomes of conventional simple missing value imputation approaches before classification to address the problem of incomplete information classification effectively. It combines the advantages of various imputation methods along with ensemble techniques. The experimental outcomes show that the suggested framework improves classification accuracy and reduces the error rate in forecasting air pollution levels.

This section explores an ensemble-imputation-classification framework that is mainly based on the following familiar imputation techniques: Single Imputation (SI), KNN Imputation (KNNI), Miss-Forest Imputation (MFI), Multiple Imputation (MI), conventional classification algorithms: Logistic Regression (LR), Naïve Bayes (NB), Support Vector Machine (SVM), K Nearest Neighbour (KNN) and most popular ensemble classification algorithms: Random Forest (RF), Bagging (Bag), Gradient Boosting (G Boost).

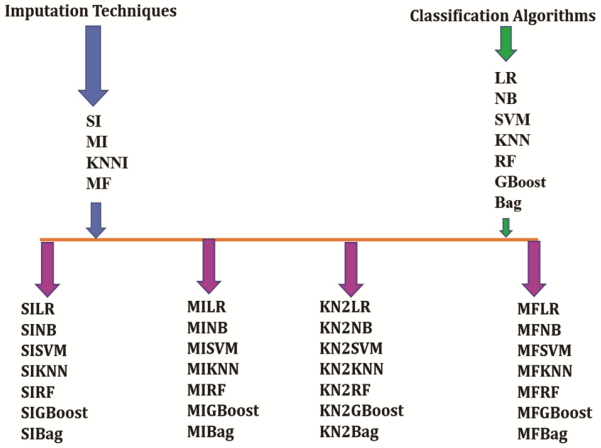

Fig. 1 — Imputation-Classification Algorithm series

Our proposed framework introduces various combinations of methodologies using several imputations, classification, and ensemble techniques, and the series is depicted in Fig. 1. All the algorithms listed in the series are applied to the outdoor air quality dataset of Beijing, China, and a brief description about the dataset is provided in the results and discussion section. The projected framework not only adventures the profits of imputation but can also be combined with ensemble techniques to improve the classification algorithm performance.

The considered air quality dataset contains more missing values in various feature columns related to air pollution concentration of SO2, NO2, PM2.5, PM10, CO, O3, and weather parameters like temperature, pressure, wind direction, and speed, rain, humidity. The direction of the wind is shown in Fig. 1. The number of missing records in the Beijing multisite air quality dataset is indicated in Table 1. The architecture of the proposed framework model is demonstrated in Fig. 2. Four imputation techniques are introduced for the treatment of missing values in the data records. To compare the achievement of each imputation technique, four copies of the dataset are taken for the experimental work.

The missing values in the first copy are imputed using Single Imputation, the second copy's missing values are imputed using Multiple Imputation, the third copy of the dataset is imputed using K Nearest Neighbour imputation and the final fourth copy is imputed using Miss-Forest Imputation respectively. Once the air pollution contaminants are imputed, individual AQI for each pollutant is calculated using the formula indicated in Eq. (1). The breakpoint values of every pollutant are prescribed by EPA – Environmental Protection Agency and listed in Table 3.

Table 1 — Dataset Statistics

	Count	Min	Max	Mean	SD	Missing (%)
PM2.5	34139	3	898	82.77	82.16	2.638
PM10	34346	2	984	110.06	95.22	2.047
SO2	34219	0.286	341	17.38	22.82	2.666
NO2	34041	2	290	59.31	37.12	2.917
CO	33288	100	10000	1262.95	1221.44	5.065
O3	33345	0.214	423	56.35	57.92	4.902
TEMP	35044	−16.8	40.5	13.58	11.39	0.057
PRES	35044	985.9	1042	1011.85	10.40	0.057
DEWP	35044	−35.3	28.5	3.12	13.69	0.057
RAIN	35044	0	72.5	0.07	0.91	0.057
WSPM	35050	0	11.2	1.71	1.20	0.231

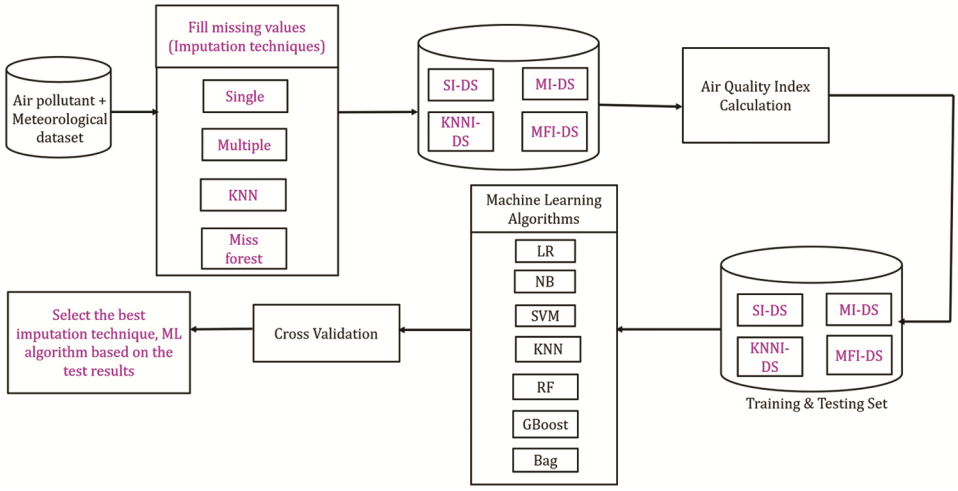

Fig. 2 — Architecture of the proposed framework

Table 2 — Air Quality Level based on AQI range (https://en.wikipedia.org/wiki/Air_quality_index)

Air Quality Index value	Air Quality Level	Class Label	Recommendations
0 – 50	Excellent	0	People continue outdoor activities
51 – 100	Good	1	Hypersensitive people should cut down on outdoor activities
101 –150	Lightly Polluted	2	Children, seniors and individuals with heart or respiratory problems trim outdoor activities
151 – 200	Moderately Polluted	3	General population should moderately reduce outdoor activities
201 – 300	Heavily Polluted	4	Stay indoors and avoid outdoor activities
>300	Severely Polluted	5	Stay indoors and avoid outdoor activities

Table 3 — Air pollutant breakpoint table by EPA (https://en.wikipedia.org/wiki/Air_quality_index)

O_3 (ppb)	O_3 (ppb)	$PM_{2.5}$ ($\mu g/m^3$)	PM_{10} ($\mu g/m^3$)	CO (ppm)	SO_2 (ppb)	NO_2 (ppb)	AQI
P_{low}–P_{high} (8 hr)	P_{low}–P_{high} (1 hr)	P_{low}–P_{high} (24 hr)	P_{low}–P_{high} (24 hr)	P_{low}–P_{high} (8 hr)	P_{low}–P_{high} (1 hr)	P_{low}–P_{high} (1 hr)	PB_{low}–PB_{high}
0 - 54	—	0.0 –12.0	0 – 54	0.0 – 4.0	0 – 35	0 – 53	0 – 50
55 – 70	—	12.1 – 35.4	55 – 154	4.5 – 9.4	36 – 75	54 – 100	51 – 100
71 – 85	125 – 164	35.5 – 55.4	155 – 254	9.5 – 12.4	76 – 185	101 – 360	101 – 150
86 – 105	165 – 204	55.5 –150.4	255 – 354	12.5 – 15.4	186 – 304	361 – 649	151 – 200
106 – 200	205 – 404	150.5 – 250.4	355–424	15.5 – 30.4	305 – 604	650 – 1249	201 – 300
—	405 – 504	250.5 – 350.4	425–504	30.5 – 40.4	605 – 804	1250 – 1649	301 – 400
—	505 – 604	350.5 – 500.4	505 – 604	40.5 – 50.4	805 – 1004	1650 – 2049	401 – 500

$$I_{AQI}=\frac{PB_{High}-PB_{Low}}{P_{High}-P_{Low}}\left(P-P_{Low}\right)+PB_{Low} \quad \ldots (1)$$

where, I_{AQI} is the Air Quality Index of Individual pollutant, PB_{High} is the breakpoint concentration to P_{High}, PB_{Low} is the breakpoint concentration to P_{Low}, P_{High} is the concentration breakpoint $\geq$ P, P_{Low} is the concentration breakpoint $\leq$ P, the value of pollutant concentration recorded in monitoring station is P.

Generally, a maximum of individual air quality index MAXAQI is considered as the final one as mentioned in Eq. (2). Based on that value, the air quality level is considered for that day, and the corresponding pollutant is declared as a primary source of quality level. Air quality levels for various AQI intervals and corresponding recommendations for the people are illustrated in Table 2.

$$FINAL_{AQI}=\max\left\{AP1_{AQI}, AP2_{AQI}, AP3_{AQI}, \ldots . APN_{AQI}\right\} \quad \ldots (2)$$

where, N is the number of air pollutants.

After storing calculated AQI values in each copy of the dataset, it is separated into training and testing sets in the proportion of 70:30. Then the classifiers are

trained using the training set and values in the test set are predicted. To avoid model over fitting, cross-validation is also done, and the mean accuracy scores are taken into consideration.

Simple Imputation Methodologies

SIBag (Simple Imputation with Bagging)

Machine Learning employs various techniques to construct models and enhance their effectiveness. Ensemble learning methods aid in the improvement of classification and regression models' accuracy. Ensemble learning is an extensively used and ideal machine learning method that combines multiple distinct models, often denoted as base models, to create effective and optimal predictive models.

Bagging (Bootstrap aggregation) is an ensemble learning technique that can help machine learning algorithms to improve their performance and accuracy.[15] It is used to cope with bias-variance trade-offs and decrease a prediction model's variance.

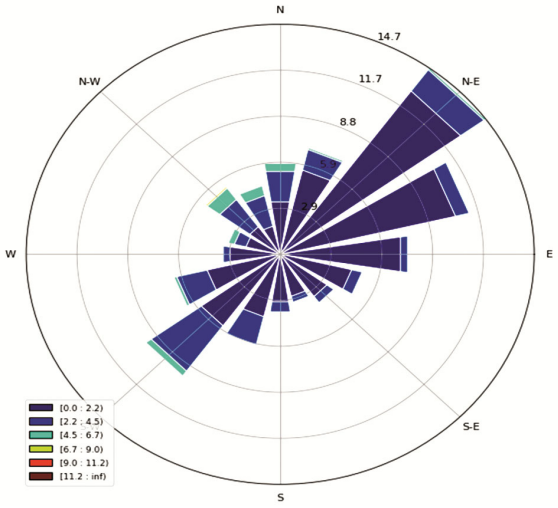

Fig. 3 — Wind direction

Bagging prevents data over fitting and is used in regression and classification models, particularly decision tree algorithms. Decision trees always suffer from high variance. A portion of the SIBag decision tree is depicted in Fig. 4. When the information samples are separated into minor parts, it is pretty to assume variance when the data sample itself is reformed. Bootstrap Aggregation of Bagging is a technique that can be used to overcome this problem. Generally, averaging a set of independent random attributes reduces variance.

Consider n random attributes with the identical variance σ^2 then the average variance is σ^2/n. This idea can be prolonged to construct multiple prediction models on diverse training data sets and the average of their outcomes to reduce the variance in the response. To train M in different models, predict the outcome using the model $f_i(x)$ and find the mean of the results to get a low variance, as shown in Eqs (3 & 4).

$$f=\frac{1}{M}\sum_{m=1}^{M} f_m(x) \quad \ldots (3)$$

$$E[f_i(x)]=\mu \; Var(f_i(x))=\sigma^2=E[f_i(x)^2]-\mu^2 \quad \ldots (4)$$

Then the variance of the average trees can be found using Eq. (5).

$$Var\left(\frac{1}{M}\sum_{m=1}^{M} f_m(x)\right)=E\left[\left(\frac{1}{M}\sum_{m=1}^{M} f_m(x)\right)^2\right]-E\left[\left(\frac{1}{M}\sum_{m=1}^{M} f_m(x)\right)\right]^2 \quad \ldots (5)$$

SIBoost (Simple Imputation with Boosting)

In this model, like the previous one, missingness is handled by mean and median, then the generated samples undergo the calculation of air quality level. The boosting technique is applied to the final record samples for training and testing. Boosting is a universal technique that can be applied to several statistical

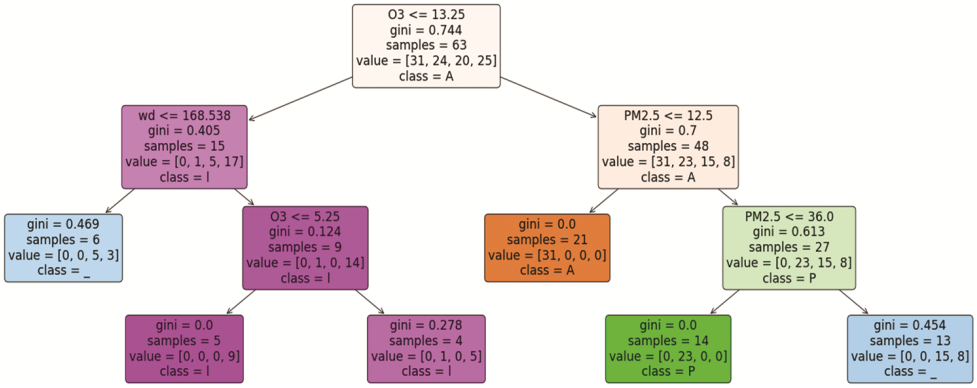

Fig. 4 — A slice of the decision tree of SIBag

learning methods. Like Bootstrapping, boosting is also trained on various data record samples generated from the training sample, but instead of constructing independent trees on different data samples, boosting is a step-by-step (sequential) procedure that develops trees on revised versions of the original records.[16] The main parameters of boosting are the number of trees NT, and the shrinkage parameter ʎ, which controls how slowly the model is learning, and it is a very small positive number. To yield better performance, a very small shrinkage parameter relies on a very large number of trees. Extreme gradient boosting optimizes the supervised learning loss function Ω using the objective function indicated in Eq. (6).

$$Obj(\theta)=NL(\theta)+\Omega(\theta) \quad \ldots (6)$$

In the above Eq. 6, the first term represents the loss function and the next one indicates the regularization term, which is used to control the model's complexity and over fitting to some level. Assume the number of trees as NT, number of leaves as NL, and number of data samples as ND. The loss function for Root Mean Square Error is shown in Eq. (7).

$$RMSE= \sum_{i=1}^{ND}(y_i-\hat{y}_i)^2 \quad \ldots (7)$$

The mathematical notation of the final prediction and the minimized objective is represented in Eqs (8 & 9).

$$\hat{y}_i= \sum_{t=1}^{NT} f_t(x_i) \quad \ldots (8)$$

$$Obj(\theta) = \sum_{i=1}^{ND} l\left(y_i, \hat{y}_i\right) + \sum_{t=1}^{NT} \Omega(f_t) \quad \ldots (9)$$

When building tree level by level, the split in each level should be obtained such that the difference between the split objectives, the current node is as huge as possible. The difference can be represented as gain as specified in Eq. (10), and the two leaves generated after a split are denoted as left (L) and right (R).

$$Gain = CurrNode_{obj} - SplitNodes_{totobj}$$

$$=\left(-\frac{1}{2}\frac{(G_L+G_R)^2}{(H_L+H_R)+ʎ}+\gamma\right)-\frac{1}{2}\frac{(G_L)^2}{(H_L)+ʎ}+\gamma+\left(-\frac{1}{2}\frac{G_R}{(H_R)+ʎ}+\gamma\right)$$

$$=\frac{1}{2}\left[\frac{G_L^2}{H_L+ʎ}+\frac{G_R}{H_R+ʎ}-\frac{(G_L+G_R)^2}{(H_L+H_R)+ʎ}\right] \quad \ldots(10)$$

Multiple Imputation with Bagging and Boosting

Multiple Imputation or Iterative imputation is a process in which each attribute is constructed as a function of the other attributes, such as a regression problem with missing values. Each characteristic is imputed one by one, allowing earlier imputed results to be used as a portion of a model to predict subsequent features. It is repetitive because this procedure is repeated several times, allowing for ever-improved forecasts of missing values to be estimated as incomplete data across all features are predicted. This method is also known as MICE - Multivariate Imputation by Chained Equations. The same raw dataset with missing records is imputed using iterative imputation, and Extreme gradient boost regression is used to predict the feature values. The workflow of Multiple Imputation is illustrated in Fig. 5. Bagging and Boosting methods are then applied to the new imputed data samples, and the comparison outcomes are deliberated in the next section.

Results and Discussion

This segment presents the detailed experimental research design, which includes details about the dataset, imputation methods, and a comparison of the results yielded by ensemble methods. Tryouts were carried out on the air quality dataset composed from various cities of Beijing, China, from the time interval March 2013 to February 2017. Yearly average statistics of each pollutant concentration are described in Fig. 6. From Fig. 7 it is ensured that there is a solid association between air pollutants and meteorological parameters.

Datasets are downloaded from the UCI repository (https://archive.ics.uci.edu/ml/datasets/ Beijing +

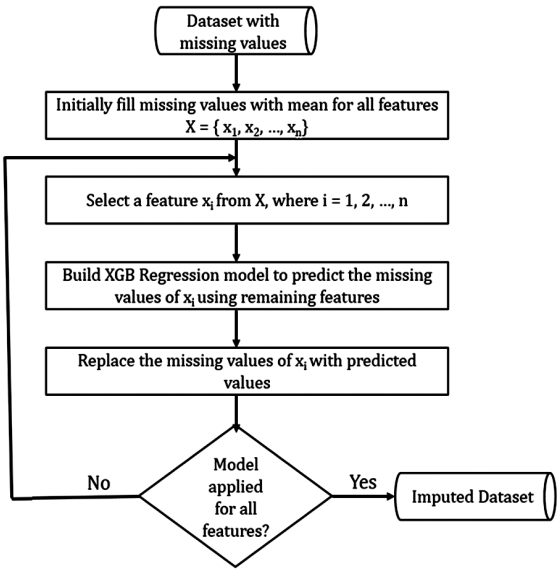

Fig. 5 — Multiple imputation workflow

Multi − Site + Air-Quality + Data).[17] Dataset consists of nearly 12 separate records for each city, and the experiments are done on a particular dataset belonging to the monitoring station Aotizhongxin. Nearly 35000 observations are recorded for that particular station in the specified time interval. Detailed information about the features of the dataset is discussed in Table 4. The dataset consists of 18 features, including air and weather information, and the correlation between each and every feature is depicted in Fig. 7.[18,19] Considered station observation consists of so many missing values, and its statistics are depicted in Table 1.

A Laptop with Intel core i5 10th Generation Quad-core processor with GeForce GTX1650 GPU, 8 GB main memory with windows 10 operating system is used to complete the experimentations. All the compilations are employed in python using ML packages. Results are compared based on the accuracy yielded by several classifier techniques and based on the efficiency of several classification techniques.

Root-mean-square error and the accuracy of the framework model are used as the evaluation metrics to measure the achievement of our suggested work. Since the proposed work is based on multi-class prediction, a balanced accuracy score is considered. K-fold cross-validation with a k value of 10 is also supported to avoid over fitting problems. Multi-class

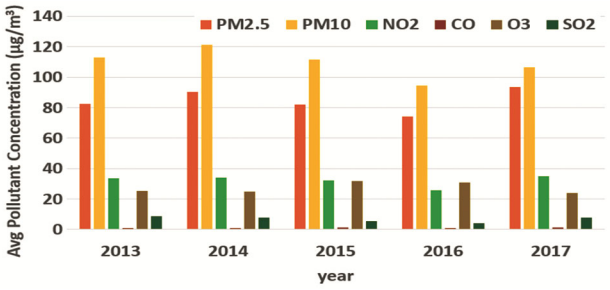

Fig. 6 — Mean Air pollution concentration for the observation interval

Table 4 — Dataset characteristics

Representation	Description
No	Row Number
Year	Year at which the Observation is recorded in this row
Day	Day at which the Observation is recorded in this row
Hour	Hour at which the Observation is recorded in this row
PM2.5	PM2.5 concentration value in µg / m^3
PM10	PM10 concentration value in µg / m^3
SO_2	SO_2 concentration value in µg / m^3
NO_2	NO_2 concentration value in µg / m^3
CO	CO concentration value in µg / m^3
O_3	O_3 concentration value in µg / m^3
TEMP	Temperature in degree Celsius
PRES	Pressure in hPa
DEWP	Dew Point temperature in degree Celsius
RAIN	Precipitation in mm
wd	Wind direction
WSPM	Wind speed in m/s
Station	Name of the air quality level monitoring station

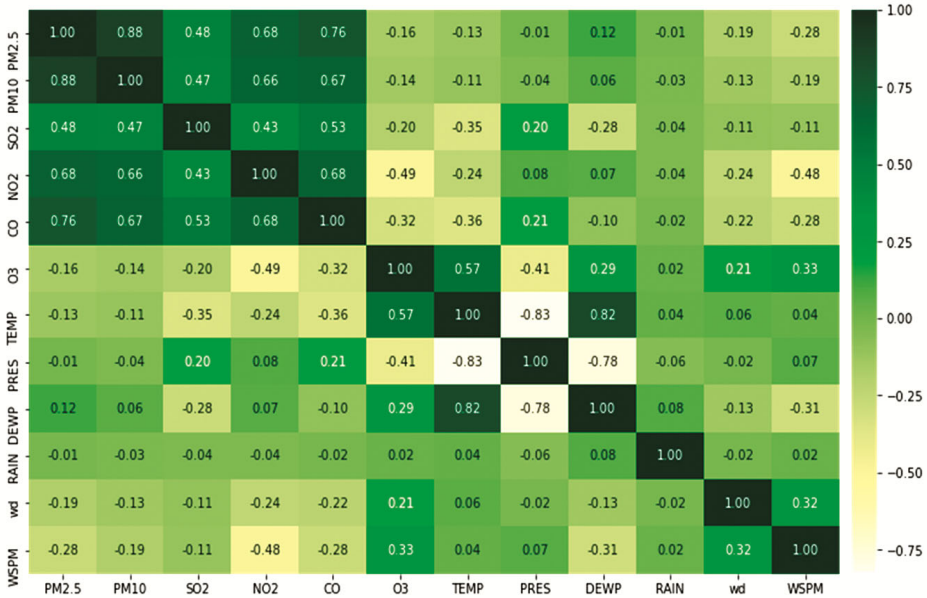

Fig. 7 — Correlation between air and weather observations

Table 5 — Accuracy score (%) and error rate of classifiers after Imputation

Classifiers	Single imputation		KNN imputation		Miss-forest imputation		Multiple imputation	
	Accuracy	Error rate	Accuracy	Error rate	Accuracy	Error rate	Accuracy	Error rate
SVM	82.76	0.473	81.73	0.465	83.65	0.477	86.89	0.465
KNN	83.28	0.446	87.47	0.438	86.18	0.442	87.53	0.431
LR	78.47	0.491	78.47	0.482	78.1	0.491	80.13	0.457
NB	80.37	0.558	83.23	0.554	83.12	0.55	83.77	0.453
RF	92.86	0.367	93.45	0.356	93.67	0.354	94.08	0.327
GBoost	95.73	0.256	95.82	0.238	95.67	0.251	97.25	0.133
Bagging	98.9	0.056	93.56	0.057	98.78	0.058	99.73	0.028

accuracy as mentioned in Eqs (11 & 12), can also be defined as the average number of correct predictions.[17] Here, NC represents the number of classifiers, and I the indicator function, which returns 1 for matching class and 0 for non-matching, Wi is the weight assigned for each class, and |S| ranges from 0 to S, which indicates the class labels.

$$\text{Accuracy}= \frac{1}{NC}\sum_{i=1}^{|S|}\sum_{j:f(x)=n} I\left(f(x)=\widehat{f(x)}\right) \quad \ldots(11)$$

$$\text{Weighted Accuracy}= \sum_{i=1}^{|S|} wi \sum_{j:f(z)=n} I\left(\left(f(x)=\widehat{f(x)}\right)\right) \quad \ldots(12)$$

Simple Imputation with Naïve Bayes suffers from a high error rate, and nearly both bagging and boosting attained a low error rate. A noteworthy issue with the whole learning algorithm is that one cannot predict how well the proposed process will perform on new information until it is tested. It can be overcome through the partitioning of record samples. The existing dataset can be partitioned into training and test sets by a 70:30 proportion. The model can be trained using a training set, prediction made on the test set, and the model is evaluated using cross-validation with 10 folds. The predicted accuracy score and error rate (RMSE) of each classifier model are shown in Table 5. Linear Regression produced less accuracy score when compared to the conventional classifiers Naïve Bayes, Support Vector Machine, and K Nearest Neighbour. Among the ensemble techniques, Bagging results are more accurate than Random Forest and Boosting. When compared to Simple, KNN, Miss-forest, multiple imputation yields better results for all classifiers.

Conclusions

Even though abundant methods exist to forecast air contamination quality, enhancements are still required in terms of performance metrics. Most approaches do not emphasize the nature of existing data and its natural influence on the final output. The proposed work ensures an increase in the accuracy of the air pollution forecasting system by utilizing the machine learning method based on imputation techniques to handle incomplete information. Further, additional attention to gathering the distinctive superiorities of various classifiers to advance the efficiency of incomplete data classification is also made. The scheme described in this work includes a novel ensemble-imputation-classification framework that attempts to revise the outcomes attained by several conventional classification algorithms. Experimental outcomes of both RMSE and average accuracy score established the dominance of the suggested framework model, and also ensemble learning on the whole data records provides robust forecasting. The proposed framework provides training using ensemble mechanisms on a collection of complete wide-ranging data samples gained from several imputation methods to handle missing data discretely. Results show that our multiple imputation yields an attractive accuracy score and RMSE rate for all classifiers when compared to other single imputation techniques.

Acknowledgement

The study is supported by FIST grant received from Department of Science and Technology, Government of India (Reference No.:SR/FST/MSI-107/2015(c)).

Conflict of Interest(COI)

The authors declare that we have no conflict of interest to report regarding the present study.

References

1 Hu J, Changes in air pollutants during the COVID-19 lockdown in Beijing: Insights from a machine-learning technique and implications for future control policy, *Atmos Ocean Sci* Lett, **14(4)** (2021) 100060.

2 Liu Y & Gopalakrishnan V, An overview and evaluation of recent machine learning imputation methods using cardiac imaging data, *Data*, **2(1)** (2017).

3 Zhao L, Liang H R, Chen F Y, Chen Z, Guan W J & Li J H, Association between air pollution and cardiovascular

mortality in China: a systematic review and meta-analysis, *Oncotarget*, **8(39)** (2017) 66438–48.

4 Pollice A & Lasinio G J, Two approaches to imputation and adjustment of air quality data from a composite monitoring network, *J Data Sci*, **7(1)** (2021) 43–59.

5 Sabarivani A, Ramadevi R, Pandian R & Krishnamoorthy N R, Effect of data preprocessing in the detection of epilepsy using machine learning techniques, *J Sci Ind Res*, **80(12)** (2021) 1066–1077.

6 Tran C T, Zhang M, Andreae P & Xue B, Multiple imputation and genetic programming for classification with incomplete data, *GECCO 2017 - Proc 2017 Genet Evol Comput Conf* 2017, 521–528.

7 Zainuri N A, Jemain A A & Muda N, A comparison of various imputation methods for missing values in air quality data, *Sains Malaysiana*, **44(3)** (2015) 449–456.

8 Yan Y, Wu Y, Du X & Zhang Y, Incomplete data ensemble classification using imputation-revision framework with local spatial neighborhood information, *Appl Soft Comput*, **99** (2021) 106905.

9 Aleryani A, Wang W & Iglesia B, Multiple imputation ensembles (MIE) for dealing with missing data, *SN Comput Sci*, **1(3)** (2020) 1–20.

10 Jadhav A, Pramod D & Ramanathan K, Comparison of performance of data imputation methods for numeric dataset, *Appl Artif Intell*, **33(10)** (2019) 913–933.

11 Liu X, Wang X, Zou L, Xia J & Pang W, Spatial imputation for air pollutants data sets via low rank matrix completion algorithm, *Environ Int*, **139** (2019) 105713.

12 Pena M, Ortega P & Orellana M, A novel imputation method for missing values in air pollutant time series data, *2019 IEEE Lat Am Conf Comput Intell* (IEEE) 2019, 1–6.

13 Ma Q & Ghosh S K, EMFlow: Data imputation in latent space via em and deep flow models, *arXiv preprint arXiv:2106.04804* (2021).

14 Yuan H, Xu G, Yao Z, Jia J & Zhang Y, Imputation of missing data in time series for air pollutants using long short-term memory recurrent neural networks, in *Proc 2018 ACM Int Joint Conf 2018 Int Symp Pervas Ubiquit Comput Wear Comput* 2018, 1293–1300, https://doi.org/10.1145/3267305.3274648.

15 Khan S S, Ahmad A & Mihailidis A, Bootstrapping and multiple imputation ensemble approach for classification problems, *J Intell Fuzzy Syst*, **37(6)** (2019) 7769–7783.

16 Kainthura P & Sharma N, Machine learning techniques to predict slope failures in Uttarkashi, Uttarakhand (India), *J Sci Ind Res (India)*, **80(1)** (2021) 66–74.

17 Xu Y, Yang W & Wang J, Air quality early-warning system for cities in China, *Atmos Environ*, **148** (2017) 239–257.

18 Lorenzo J S L, Tam W W S & Seow W J, Association between air quality, meteorological factors and COVID-19 infection case numbers, *Environ Res*, **197** (2021) 111024.

19 Dutta S, Ghosh S & Dinda S, Urban air-quality assessment and inferring the association between different factors: a comparative study among Delhi, Kolkata and Chennai Megacity of India, *Aerosol Sci Eng*, **5** (2021) 93–111.