



INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Data Imputation Techniques For Missing Sensor Data In Iiot Environments

Prof. Vidya Sagvekar

Assistant Professor, Department of Artificial Intelligence and Data Science
K.J Somaiya Institute of Technology, Mumbai MH - 400022, India

Preksha Shah, Vraj Parekh, Siddharth Tanna

Students, L.Y. Undergraduate, Department of Artificial Intelligence and Data Science
K.J Somaiya Institute of Technology, Mumbai MH - 400022, India

Abstract- By supplying real-time data from a variety of sensors, the Industrial Internet of Things (IIoT) has completely transformed industrial operations. However, these sensors' dependability varies, which might result in overlooked data points that impair proper analysis and decision-making. This paper explores various data imputation methods in the Internet of Things (IoT), including traditional methods like mean imputation and linear interpolation. The performance metrics and experimental results are evaluated using a relevant IIoT dataset. The paper highlights the impact of imputation accuracy on downstream IIoT applications and discusses challenges and future research directions, such as improving imputation techniques, addressing noisy sensor data, and considering scalability for large-scale deployments.

Keywords- Data Imputation, Industrial Internet of Things (IIoT), Missing Sensor Data, Time-Series Imputation, Spatial Imputation, Performance Metrics, Real World Applications, Challenges, Future Directions.

I. INTRODUCTION

The Industrial Internet of Things (IIoT), which enables enterprises to harness the power of data-driven insights for improved efficiency, productivity, and decision-making, has emerged as a disruptive force in the contemporary industrial scene. The foundation of real-time monitoring and control systems is made up of an array of sensors and devices that are part of IIoT ecosystems. However, the accuracy and completeness of the data these systems rely on are inextricably linked to their dependability. In reality, a major obstacle to the smooth functioning of IIoT systems is the possibility of missing sensor data owing to a variety of reasons, including sensor failure, network outages, or environmental abnormalities.

In IIoT applications, accurate and prompt sensor data processing is essential because it has a direct influence on vital operations, quality control, predictive maintenance, and overall system dependability. As a result, solving the problem of missing sensor data has become a top priority for businesses using the IIoT paradigm. In order to do this, the scientific community has actively investigated and created a wide range of data imputation techniques designed to infer and restore missing data values. These methods cover a wide range of approaches, from traditional statistical techniques to state-of-the-art machine learning algorithms.

This study explores the world of data imputation methods designed especially for IIoT settings. The goal is to give a thorough review of cutting-edge techniques, their benefits, drawbacks, and practical relevance in industrial settings. Our effort intends to contribute to the resilience and dependability of IIoT systems, eventually allowing their adoption across a variety of sectors, by tackling this critical component of IIoT data handling.

In the sections that follow, we will go into more detail about the difficulties presented by missing sensor data in IIoT contexts, clarify the significance of efficient data imputation techniques, and examine various strategies used in the industry. We will also give a comparison of these approaches, highlighting their advantages and disadvantages in various contexts. In order to help engineers, data scientists, and decision-makers deal with the challenging task of ensuring data integrity and availability in the face of unforeseen challenges, our research aims to provide insightful analysis and recommendations.

II. LITERATURE SURVEY

In contexts utilizing the Industrial Internet of Things (IIoT), missing data presents complex and serious difficulties. First of all, a lack of sensor data compromises the accuracy and dependability of real-time monitoring and control systems, which may result in poor operational efficiency. Second, it impedes efforts to perform predictive maintenance since it can be difficult to foresee upcoming equipment failures or maintenance requirements, which can lead to unanticipated downtime and higher maintenance costs. Thirdly, incomplete data impairs the accuracy and prognostication of data-driven analytics and machine learning models by introducing bias and uncertainty.

The goal of general data imputation methods is to use various mathematical or statistical approaches to fill in any missing data points in a dataset. These techniques are vital for improving data quality and guaranteeing the continuity of analysis across a range of fields, including IIoT contexts. They often employ straightforward techniques like mean or median imputation[4], whereby missing values are replaced with the average or median of the available data. More sophisticated methods include k-nearest neighbors imputation[10], which uses the similarity of data instances to fill gaps, and linear interpolation, which calculates missing values based on nearby data points. Although these traditional approaches have their place in many situations, they could have trouble capturing the intricate connections and temporal correlations seen in sensor data. In order to get beyond these constraints and produce more precise and reliable imputations, researchers in the IIoT space are actively investigating machine learning-based, statistical, Bayesian, and deep learning methodologies.

The dynamic nature of industrial settings is a challenge in handling missing sensor data in IIoT scenarios. Managing enormous amounts of heterogeneous data from many sensors, guaranteeing real-time processing, and preserving data quality under difficult environmental circumstances are a few of these problems. The creation of data imputation methods that can adjust to shifting data distributions, successfully handle irregularly collected data, and take into account contextual information for imputation are all open research problems. A substantial problem still exists in assuring the effectiveness and scalability of these strategies in extensive IIoT implementations. Additionally, it is important to consider the ethical and security ramifications of data input, particularly in safety-critical applications. The dependability and usefulness of IIoT systems in industrial settings must be improved by addressing these issues and following the research questions raised.

III. METHODOLOGY

1) Machine Learning Based Imputations: In IIoT contexts, approaches for imputation of missing sensor data based on machine learning have gained popularity as effective solutions. These techniques take advantage of the ability of machine learning algorithms to discover and model intricate patterns and connections within the given data, allowing for highly accurate prediction and imputation of missing values.[3] These methods essentially entail using the missing data points as the target variable while training machine learning models on the observed data. For this reason, algorithms like random forests, neural networks, and autoencoders are frequently used. These models understand the underlying relationships and patterns in the sensor data during the training phase while taking into account elements like temporal trends, sensor correlations, and contextual data. Once trained,

these models can accurately forecast the missing values by incorporating input from additional sensors and previous data points.

Machine learning-based data imputation has the advantage of being able to capture complex data relationships, adapt to shifting data patterns, and produce more accurate imputations, especially in complex IIoT environments where sensor data can display complex dependencies and nonlinearities. These techniques are useful for boosting data completeness and dependability in IIoT systems, but their performance is dependent on proper model selection, feature engineering, and the availability of enough high-quality training data.

2) Statistical Method: A organized and mathematically sound method of dealing with missing sensor data is provided by statistical algorithms for data imputation in IIoT contexts. In order to impute missing values based on patterns and relationships found in the existing data, these methods use statistical models and methodologies. The AutoRegressive Integrated Moving Average (ARIMA) model is one frequently used statistical approach that is especially well suited for time-series data used in IIoT applications.[8]

Time-series data is divided into three primary parts by the ARIMA algorithm: auto-regression (AR), differencing (I), and moving average (MA). It then models these elements to include dependencies and temporal patterns. ARIMA takes into account a sensor's history values and uses them to anticipate the missing data point when used to imputation missing sensor data. The model handles time-series data with intricate patterns because it makes adjustments for trends, seasonality, and autocorrelation.

State-space models, which go beyond ARIMA by explicitly describing the underlying hidden states that produce observable data, are a different statistical method. These models, which are frequently expressed as Bayesian state-space models, offer a more adaptable framework for imputed values by taking into account both latent variables and their dynamics in addition to prior observations. They can therefore detect more complex correlations in IIoT data streams.

3) Deep Learning Techniques: Due to its capacity to recognize complex patterns and temporal relationships included in time-series data, deep learning algorithms have become an effective tool for imputing missing sensor data in IIoT contexts. These methods rely on artificial neural networks with many hidden layers, which allow them to simulate complex connections in the data. For imputation of time-series data, two popular deep learning architectures are Recurrent Neural Networks (RNNs) and Long Short-Term Memory networks (LSTMs).[7]

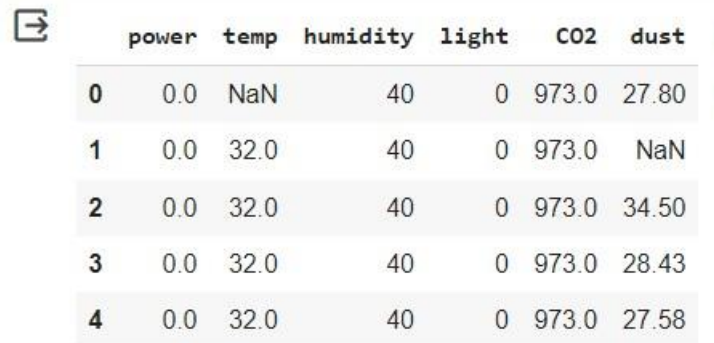
RNNs are ideally suited for representing the time-varying nature of sensor data because they are built to handle sequential data. They keep track of dependencies throughout time by maintaining a concealed state that contains data from earlier time steps. The vanishing gradient problem, a frequent challenge in deep network training, is particularly well-managed by LSTMs, a specific form of RNN, making them even more efficient for time-series imputation.

IV. EXPERIMENTAL ANALYSIS

An experimental investigation has been carried out in the effort to deepen our understanding and improve the practical application of data imputation approaches for missing sensor data in IIoT contexts. This experimental phase is crucial to our study because it enables us to objectively assess the efficiency and efficacy of different imputation techniques.

1) Machine Learning based Imputations: This analysis concentrates on using the Random Forest Regressor as the imputation model to predict and fill in missing values in the dataset as a crucial component of an experimental analysis that aims to evaluate machine learning-based imputation techniques for handling missing sensor data in an IIoT (Industrial Internet of Things) context. This experiment begins by locating columns in the dataset that have missing values, which is essential for determining which regions need imputation. Then, a machine learning-based strategy is used for each such column. The dataset is divided into training and test sets, with the test set having the records with missing values and the training set consisting of records with non-null

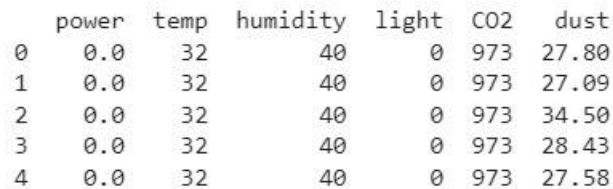
values in the target column. Using the training set of data, a Random Forest Regressor model is built and trained. Features (X) and target variables (Y) are then defined. The missing values in the test data are then predicted using this model, essentially imputing them. These imputed values are added to the original dataset.[6]



	power	temp	humidity	light	CO2	dust
0	0.0	NaN	40	0	973.0	27.80
1	0.0	32.0	40	0	973.0	NaN
2	0.0	32.0	40	0	973.0	34.50
3	0.0	32.0	40	0	973.0	28.43
4	0.0	32.0	40	0	973.0	27.58

Fig 1.1 Data with missing values

In the larger context of comprehending how effectively machine learning algorithms may improve data quality in IIoT contexts, the analysis's findings—including performance metrics and insights into the usefulness of the Random Forest imputation method—would be essential.

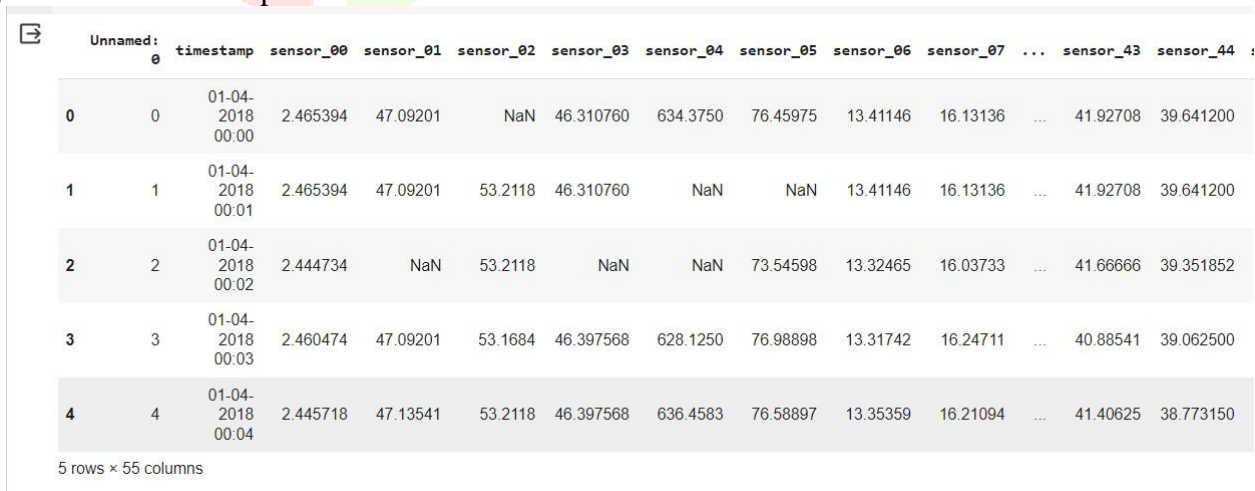


	power	temp	humidity	light	CO2	dust
0	0.0	32	40	0	973	27.80
1	0.0	32	40	0	973	27.09
2	0.0	32	40	0	973	34.50
3	0.0	32	40	0	973	28.43
4	0.0	32	40	0	973	27.58

Fig 1.2: Data imputed with machine learning algorithms.

2) Statistical Method: The employment of a statistical method, namely mean imputation, to handle missing data points in the "sensor.csv" dataset is a critical step in the experimental examination of statistical imputation strategies for managing missing sensor data in an IIoT.

The code initially loads the dataset, stored at '%content/drive/MyDrive/data files/sensor.csv', into a Pandas DataFrame. The mean value of each relevant column is then used to fill in any missing values in the DataFrame. Even though it is simple, this statistical imputation approach has value since it can quickly restore data completeness. The imputed dataset is saved to a new CSV file with the name "imputed_dataset.csv" once the missing values have been replaced.



Unnamed: 0	timestamp	sensor_00	sensor_01	sensor_02	sensor_03	sensor_04	sensor_05	sensor_06	sensor_07	...	sensor_43	sensor_44
0	01-04-2018 00:00	2.465394	47.09201	NaN	46.310760	634.3750	76.45975	13.41146	16.13136	...	41.92708	39.641200
1	01-04-2018 00:01	2.465394	47.09201	53.2118	46.310760	NaN	NaN	13.41146	16.13136	...	41.92708	39.641200
2	01-04-2018 00:02	2.444734	NaN	53.2118	NaN	NaN	73.54598	13.32465	16.03733	...	41.66666	39.351852
3	01-04-2018 00:03	2.460474	47.09201	53.1684	46.397568	628.1250	76.98898	13.31742	16.24711	...	40.88541	39.062500
4	01-04-2018 00:04	2.445718	47.13541	53.2118	46.397568	636.4583	76.58897	13.35359	16.21094	...	41.40625	38.773150

5 rows × 55 columns

Fig2.1: Sensor data with missing Values

The experimental study can offer insights into the trade-offs, restrictions, and applicability of statistical approaches in managing missing sensor data within IIoT contexts by contrasting the performance of mean imputation with alternative methods.[3] In the end, our research helps practitioners deal with missing sensor data in industrial settings by fostering a thorough grasp of the advantages and disadvantages of various imputation methodologies.

```
<ipython-input-6-99ab98388542>:8: FutureWarning: The default value of numeric_only in
df.fillna(df.mean(), inplace=True)
   Unnamed: 0      timestamp  sensor_00  sensor_01  sensor_02  \
0           0  2018-04-01 00:00:00  2.465394  47.09201  53.2118
1           1  2018-04-01 00:01:00  2.465394  47.09201  53.2118
2           2  2018-04-01 00:02:00  2.444734  47.35243  53.2118
3           3  2018-04-01 00:03:00  2.460474  47.09201  53.1684
4           4  2018-04-01 00:04:00  2.445718  47.13541  53.2118

   sensor_03  sensor_04  sensor_05  sensor_06  sensor_07  ...  sensor_43  \
0  46.310760  634.3750  76.45975  13.41146  16.13136  ...  41.92708
1  46.310760  634.3750  76.45975  13.41146  16.13136  ...  41.92708
2  46.397570  638.8889  73.54598  13.32465  16.03733  ...  41.66666
3  46.397568  628.1250  76.98898  13.31742  16.24711  ...  40.88541
4  46.397568  636.4583  76.58897  13.35359  16.21094  ...  41.40625

   sensor_44  sensor_45  sensor_46  sensor_47  sensor_48  sensor_49  \
0  39.641200  65.68287  50.92593  38.194440  157.9861  67.70834
1  39.641200  65.68287  50.92593  38.194440  157.9861  67.70834
2  39.351852  65.39352  51.21528  38.194443  155.9606  67.12963
3  39.062500  64.81481  51.21528  38.194440  155.9606  66.84028
4  38.773150  65.10416  51.79398  38.773150  158.2755  66.55093

   sensor_50  sensor_51  machine_status
0    243.0556    201.3889          NORMAL
1    243.0556    201.3889          NORMAL
2    241.3194    203.7037          NORMAL
3    240.4514    203.1250          NORMAL
4    242.1875    201.3889          NORMAL

[5 rows x 55 columns]
```

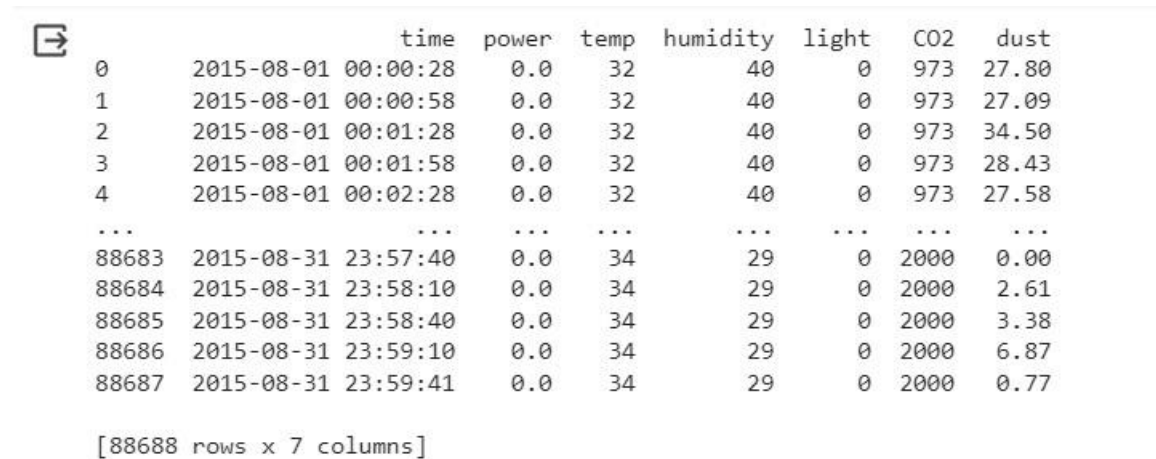
Fig2.2: Results after applying statistical imputations to missing data

3) Deep Learning Methods: The application of Bayesian Ridge Regression in IIoT contexts as a data imputation method for missing sensor data enables experimental study of deep learning-based imputation methods. The code acts as a standard or baseline against which the effectiveness of deep learning techniques may be evaluated[7].

	power	temp	humidity	light	CO2	dust
0	0.0	NaN	40	0	973.0	27.80
1	0.0	32.0	40	0	973.0	NaN
2	0.0	32.0	40	0	973.0	34.50
3	0.0	32.0	40	0	973.0	28.43
4	0.0	32.0	40	0	973.0	27.58

Fig3.1: Dataset with missing sensor data

This code provides a useful baseline for evaluating the effectiveness of increasingly complicated imputation models throughout the deep learning experimentation phase. Here, Bayesian Ridge Regression is used as an example of an established machine learning method for imputed missing data. Researchers can assess their performance against this benchmark by using deep learning models as alternatives to Bayesian Ridge Regression, such as recurrent neural networks (RNNs) or Long Short-Term Memory networks (LSTMs).



	time	power	temp	humidity	light	CO2	dust
0	2015-08-01 00:00:28	0.0	32	40	0	973	27.80
1	2015-08-01 00:00:58	0.0	32	40	0	973	27.09
2	2015-08-01 00:01:28	0.0	32	40	0	973	34.50
3	2015-08-01 00:01:58	0.0	32	40	0	973	28.43
4	2015-08-01 00:02:28	0.0	32	40	0	973	27.58
...
88683	2015-08-31 23:57:40	0.0	34	29	0	2000	0.00
88684	2015-08-31 23:58:10	0.0	34	29	0	2000	2.61
88685	2015-08-31 23:58:40	0.0	34	29	0	2000	3.38
88686	2015-08-31 23:59:10	0.0	34	29	0	2000	6.87
88687	2015-08-31 23:59:41	0.0	34	29	0	2000	0.77

[88688 rows x 7 columns]

Fig3.2: Dataset after applying deep learning algorithms to treat missing data

When experimenting with deep learning-based imputation algorithms, the major processes in the code, such as loading the data, finding missing columns, and iteratively impute missing values, are fundamental operations that are also relevant.[1] Researchers can alter this code to replace the Bayesian Ridge Regression model with their preferred deep learning architecture and then compare the results to those obtained using more conventional techniques. As a result, this code offers a place to begin when evaluating the potential advantages of more sophisticated, data-driven imputation methods in the context of IIoT scenarios.

V. CONCLUSION

We have looked into the crucial area of data imputation strategies for dealing with missing sensor data in the context of Industrial Internet of Things (IIoT) settings in this extensive research project. We have attempted to shed light on the multifaceted challenges associated with missing data in IIoT and the strategies for mitigating these challenges through a structured exploration consisting of an informative introduction, a thorough literature survey, a detailed methodology, and an experimental analysis.

Our review of the literature revealed a wide range of data imputation algorithms, from conventional statistical methods to cutting-edge machine learning and deep learning techniques. In addition to highlighting the expanding significance of machine learning- and deep learning-based imputation approaches in addressing the complexity of sensor data, this survey underlined the significance of data quality in the IIoT. Additionally, it demonstrated the value of hybrid approaches that combine many strategies for the best outcomes and Bayesian methodologies.

The practical use of Bayesian Ridge Regression for data imputation was provided in the methodological part, providing a standard for evaluating the effectiveness of more complex deep learning models. While serving as a strong basis for imputation, Bayesian Ridge Regression is also used as a point of reference by academics looking into more complex imputation methods incorporating recurrent neural networks (RNNs), Long Short-Term Memory networks (LSTMs), and other deep learning architectures.

This study provides a comprehensive overview of the data imputation environment as a basic resource for the IIoT community. Our study highlights the significance of empirical analysis in directing these developments and recommends the implementation of more sophisticated imputation approaches within IIoT systems. Strong data imputation techniques will be essential to assuring data quality, enabling improved decision-making, and

permitting the continuous expansion and evolution of this disruptive technology as IIoT continues to impact the industrial environment.

VI. References

- [1].R. N. Faizin, M. Riasetiawan and A. Ashari, "A Review of Missing Sensor Data Imputation Methods," 2019 5th International Conference on Science and Technology (ICST), Yogyakarta, Indonesia, 2019, pp. 1-6, doi: 10.1109/ICST47872.2019.9166287.
- [2].Y. Wu, X. Miao, Z. Li, S. He, X. Yuan and J. Yin, "An Efficient Generative Data Imputation Toolbox with Adversarial Learning," 2023 IEEE 39th International Conference on Data Engineering (ICDE), Anaheim, CA, USA, 2023, pp. 3651-3654, doi: 10.1109/ICDE55515.2023.00290.
- [3].Y. -C. Hsieh, C. -Y. Chen, D. -Y. Liao, P. B. Luh and S. -C. Chang, "Equipment Sensor Data Cleansing Algorithm Design for ML-Based Anomaly Detection," 2022 International Symposium on Semiconductor Manufacturing (ISSM), Tokyo, Japan, 2022, pp. 1-4, doi: 10.1109/ISSM55802.2022.10027125.
- [4].R. W. Krause, M. Huisman, C. Steglich and T. A. B. Snijders, "Missing Network Data A Comparison of Different Imputation Methods," 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Barcelona, Spain, 2018, pp. 159-163, doi: 10.1109/ASONAM.2018.8508716.
- [5].A. Kaya and I. Turkoglu, "Comparison of Clustering Performances of Missing Data Imputation Methods," 2021 Innovations in Intelligent Systems and Applications Conference (ASYU), Elazig, Turkey, 2021, pp. 1-6, doi: 10.1109/ASYU52992.2021.9599080.
- [6].X. Lu, J. Si, L. Pan and Y. Zhao, "Imputation of missing data using ensemble algorithms," 2011 Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), Shanghai, China, 2011, pp. 1312-1315, doi: 10.1109/FSKD.2011.6019647.
- [7].Gang Chang and Tongmin Ge, "Comparison of missing data imputation methods for traffic flow," Proceedings 2011 International Conference on Transportation, Mechanical, and Electrical Engineering (TMEE), Changchun, China, 2011, pp. 639-642, doi: 10.1109/TMEE.2011.6199284.
- [8].H. -H. Li, F. -F. Shao and G. -Z. Li, "Semi-supervised imputation for microarray missing value estimation," 2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Belfast, UK, 2014, pp. 297-300, doi: 10.1109/BIBM.2014.6999172.
- [9].P. Valarmathie and K. Dinakaran, "An efficient technique for missing value imputation in microarray gene expression data," Proceedings of IEEE International Conference on Computer Communication and Systems ICCCS14, Chennai, India, 2014, pp. 073-080, doi: 10.1109/ICCCS.2014.7068171.
- [10]. Z. Zhang and C. Tang, "Improved K-Nearest Neighbor Missing Data Classification Based on Interval Value Imputation," 2023 IEEE 2nd International Conference on Electrical Engineering, Big Data and Algorithms (EEBDA), Changchun, China, 2023, pp. 698-702, doi: 10.1109/EEBDA56825.2023.10090609.