A Robust Framework for Journal Impact Factor Prediction Using Machine Learning and Missing Data Imputation

K.Pavani[1], ,M.Lakshmi Durga[2]

#1 Assistant Professor & Head of Department of MCA, SRK Institute of Technology, Vijayawada.

#2 Student in the Department of MCA, SRK Institute of Technology, Vijayawada

Abstract: Forecasting the Journal Impact Factor (JIF) is essential for evaluating journal quality, but missing data in real-world datasets reduces prediction accuracy. This paper proposes a machine learning-based approach that utilizes K-Nearest Neighbor Imputation (KNNI) to effectively handle missing values. The imputed dataset is then processed using a Support Vector Machine (SVM), which is further optimized by tuning key parameters such as regularization and iteration limits. Experimental results demonstrate that the optimized SVM with KNNI significantly outperforms traditional imputation methods like mean substitution, achieving lower RMSE and MAE values. The proposed model provides accurate and reliable JIF predictions, supporting better decision-making in academic publishing.

Index terms - *Journal Impact Factor (JIF), K-Nearest Neighbor Imputation (KNNI), Support Vector Machine (SVM), Missing Data Handling, Machine Learning, Prediction Model, RMSE, MAE*

1. INTRODUCTION

The Journal Impact Factor (JIF) is a widely recognized metric used to evaluate the significance and influence of academic journals based on citation analysis. It reflects the average number of citations received by articles published in a journal over a specific period, typically two years. Researchers, institutions, and funding agencies rely on JIF to assess research quality, rank journals, and make informed publication decisions. However, JIF varies across disciplines and may be influenced by highly cited papers, making accurate prediction a challenging task.

One of the major issues in forecasting JIF is the presence of missing or incomplete data in journal datasets. Traditional imputation techniques such as mean and median replacement often fail to capture the underlying relationships between data points, leading to reduced prediction accuracy. To address this limitation, this paper proposes a machine learning-based approach that integrates K-Nearest Neighbor Imputation (KNNI) with an optimized Support Vector Machine (SVM) model. The KNNI method effectively handles missing values by utilizing similarity between data instances, while the optimized SVM enhances prediction performance through parameter tuning. The proposed system aims to provide accurate and reliable forecasting of Journal Impact Factor, thereby assisting researchers in selecting appropriate journals and improving decision-making in academic publishing.

2. LITERATURE SURVEY

2.1 Analysis of Effects to Journal Impact Factors Based on Citation Networks Generated via Social Computing:

A basic experimental model is created using the concept of social computing as a guide to investigate how journal impact factors vary, especially across fields. Our technique aims to replicate, in a distributed fashion, the publishing and citation patterns of publications in journals within a related field. The process of effect from several basic components to the trend of impact factor is better exposed based on the citation networks formed. The average number of references, the average review cycle, and the annual distribution of references are some of these variables. Additionally, the examination shows that the model's capacity to approximate real data and simulation outcomes is excellent.

2.2 Dynamics of Journal Impact Factors:

Increasing a journal's impact factor is a key goal for a journal management. The impact factor is a measurement of how frequently a journal's average article is referenced over a specific time frame. The interaction of several concepts, including seniority of authors and reviewers, journal regulations, online availability of journals, and contribution quality, results in the dynamics of impact factors. The goal of this research is to explore three methods for improving a journal's impact factor and underlying resources in a sustainable manner. In order to do this, a structural simulation model that is utilized for strategy trials captures the resources and assets of a publication. Three insights are provided by the paper: It stresses the necessity of developing the levels and growth rates of the essential resources, authors and reviewers, in dynamic correspondence and offers a dynamic hypothesis about the causal structures behind a journal impact factor. Lastly, building up the pool of excellent reviewers takes time and money, but it is more stable than the pool of writers, thus it has a greater chance of leading the journal into a sustainable development regime. Future directions and limitations are spoken about.

2.3 Analysis of Machine Learning Based Imputation of Missing Data:

The availability of missing data in datasets can have an impact on data analysis and categorization. Either deletion-based or imputation-based techniques are employed to deal with missing data, which either reduce the number of data records or impute an incorrect anticipated value. If machine learning methods are used to properly produce missing values, the quality of imputed data can be greatly enhanced. This article analyzes machine learning-based imputation techniques for missing data. Machine learning is used to impute missing values in datasets using the K-nearest neighbors (KNN) and Sequential KNN (SKNN) methods. To assess the efficacy of imputed data, missing values handled by statistical deletion techniques (List-wise Deletion (LD)) and ML-based imputation techniques (KNN and SKNN) are investigated and compared using several ML classifiers (Support Vector Machine and Decision Tree). The accuracy of the used methods is compared, and the findings show that the ML-based imputation technique (SKNN) is more successful at managing missing data in nearly all datasets using both classification algorithms (SVM and DT) than the LD-based approach and KNN method.

2.4 Machine-Learning-Based Imputation Method for Filling Missing Values in Ground Meteorological Observation Data:

Ground meteorological observation data (GMOD) is the foundation of earth-related research and a crucial resource for life and social production. Unfortunately, missing values in GMOD might happen as a result of equipment malfunctions or operational problems. As a result, imputation of missing data is a common problem in GMOD pre-processing. Few studies address imputation when multiple elements are randomly missing in the dataset, despite the fact that many machine-learning techniques have been applied to the field of meteorological missing value imputation and have produced positive results. In order to impute the GMOD with random missing values in multiple attributes, this paper developed a machine-learning-based multidimensional meteorological data imputation framework (MMDIF). Based on the MMDIF, 20 machine-learning methods were tested for their efficacy in imputing missing values in 124 meteorological stations across six different climatic regions. The findings demonstrate that MMDIF-RF was the most successful missing value imputation technique, outperforming other approaches for imputing eleven different kinds of hourly meteorological variables. While MMDIF was used in this study to impute missing values in meteorological data, the technique can potentially offer recommendations for dataset reconstruction in other sectors.

2.5 Missing value imputation in multivariate time series with end-to-end generative adversarial networks:

The effectiveness of follow-up analytic applications on the multivariate time series is negatively impacted by missing values, which are present in multivariate time series for a variety of causes, including collecting mistakes. To lessen the impact of missing values on multivariate time series analysis, a variety of missing value imputation techniques have been put forth. The GAN-2-Stage has recently been utilized to solve the imputation problem with the generative model, motivated by the success of generative adversarial networks (GANs) in picture production. In particular, GANs are used by GAN-2-Stage to infer the missing data. Nevertheless, an additional stage is needed to improve the generator's input random "noise." Additionally, because to the unstable generation process and the challenge of training a GAN, the imputed values may deviate significantly from actual values. In order to impute the missing values in a multivariate time series, this research suggests an end-to-end model. In particular, we remove the input optimization stage in the GAN-2-Stage by including an encoder network into the conventional GAN design. In order to make the imputed values near to the genuine values, our generator uses real data during training. The suggested model performs better than state-of-the-art techniques in imputation tasks and downstream applications, such as regression and classification, according to experiments conducted on three real-world multivariate time series datasets.

3. METHODOLOGY

i) Proposed Work:

The proposed work focuses on developing an efficient machine learning framework to forecast the

Journal Impact Factor (JIF) in the presence of missing data. Initially, the dataset is preprocessed by handling categorical values through label encoding and identifying missing entries. To overcome the limitations of traditional imputation techniques, a K-Nearest Neighbor Imputation (KNNI) method is applied, which replaces missing values based on the similarity between data points, thereby preserving the inherent relationships within the dataset.

After imputation, the data is normalized and divided into training and testing sets for model development. A Support Vector Machine (SVM) algorithm is then employed for JIF prediction, and its performance is further enhanced by tuning key parameters such as the regularization parameter (C), maximum iterations, and cache size. The proposed model is evaluated using performance metrics like Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). The results demonstrate that the optimized SVM combined with KNNI provides more accurate and reliable predictions compared to traditional approaches, making it suitable for real-world journal evaluation tasks.

ii) System Architecture:

The system architecture illustrates the complete workflow for forecasting the Journal Impact Factor (JIF) using machine learning techniques. Initially, the dataset is collected and passed through the preprocessing stage, where operations such as label encoding, handling missing values, KNN-based feature selection (KNNI), and data normalization are performed. This stage ensures that the dataset is clean, consistent, and suitable for model training. After preprocessing, the dataset is divided into training and testing subsets to enable proper evaluation of the model.

In the next phase, multiple models are developed, including the existing SVM with mean imputation, the proposed SVM with KNN imputation, and the optimized SVM with KNN imputation. These models are trained using the processed training data and then evaluated using the test data. The trained model generates predictions for the Journal Impact Factor, and its performance is measured using evaluation metrics such as Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). The architecture ensures a systematic flow from data input to final prediction, highlighting the effectiveness of KNN-based imputation and SVM optimization in improving prediction accuracy.

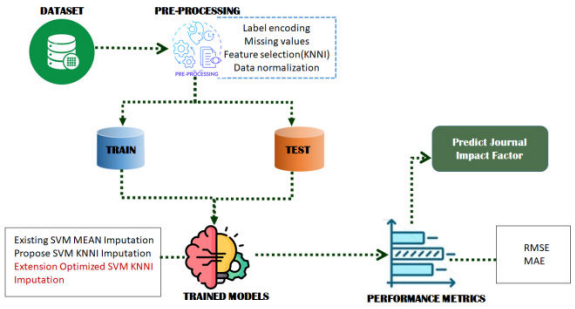

Fig.1. Proposed Architecture

iii) MODULES:

1. Data Collection Module

This module is responsible for gathering the Journal Impact Factor dataset from sources like Kaggle. It provides the raw input data required for training and testing the model.

2. Data Preprocessing Module

In this module, categorical data is converted into numerical form using label encoding, and missing

values are identified. It prepares the dataset for further processing.

3. Imputation Module (KNNI)

This module handles missing data using K-Nearest Neighbor Imputation, where missing values are replaced based on similar data points. It improves data quality compared to traditional methods.

4. Data Normalization Module

Normalization scales the dataset values into a standard range, ensuring that all features contribute equally to the model performance and improving convergence speed.

5. Data Splitting Module

The processed dataset is divided into training and testing sets. This helps in evaluating the model's performance on unseen data.

6. Model Training Module

Different models are trained in this phase, including SVM with mean imputation, SVM with KNN imputation, and optimized SVM with KNNI. It builds the prediction models.

7. Prediction Module

This module uses the trained model to predict the Journal Impact Factor for new or test data inputs.

8. Performance Evaluation Module

The system evaluates model accuracy using metrics such as RMSE and MAE, helping to compare different approaches and identify the best-performing model.

iv) ALGORITHMS:

1. Support Vector Machine (SVM) with Mean Imputation

In this approach, missing values in the dataset are replaced using mean imputation, where each missing value is substituted with the average of its respective feature. The processed data is then fed into the Support Vector Machine (SVM), a supervised learning algorithm that constructs an optimal hyperplane for regression. However, since mean imputation ignores relationships between data points, it may reduce prediction accuracy.

2. Support Vector Machine (SVM) with KNN Imputation (KNNI)

In this method, missing values are handled using K-Nearest Neighbor Imputation, where each missing value is estimated based on the values of its nearest neighbors. This preserves the inherent structure and relationships in the dataset. The imputed data is then used to train the SVM model, resulting in improved prediction accuracy compared to mean imputation.

3. Optimized SVM with KNN Imputation

This is the proposed and enhanced approach where SVM is combined with KNN-based imputation and further optimized by tuning hyperparameters such as the regularization parameter (C), maximum iterations, and cache size. This optimization improves the model's generalization ability and reduces prediction errors. As a result, this method achieves lower RMSE and MAE values, making it the most accurate and reliable algorithm for forecasting Journal Impact Factor.

4. EXPERIMENTAL RESULTS

The experimental evaluation was conducted using the Journal Impact Factor dataset after applying preprocessing techniques such as label encoding,

missing value handling, and normalization. The performance of three models—SVM with Mean Imputation, SVM with KNN Imputation (KNNI), and Optimized SVM with KNNI—was analyzed using evaluation metrics like Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). These metrics help measure the difference between actual and predicted values, where lower values indicate better model performance.

The results show that the SVM model with mean imputation produced higher error values due to poor handling of missing data. In contrast, the SVM with KNNI significantly reduced errors by preserving data relationships during imputation. The best performance was achieved by the optimized SVM with KNNI, which obtained the lowest RMSE of 0.052253 and MAE of 0.031325, outperforming all other models. Graphical analysis also confirmed that predicted values closely match actual values in the proposed approach. These results clearly demonstrate that combining KNN-based imputation with optimized SVM improves prediction accuracy and reliability for Journal Impact Factor forecasting.

Accuracy: The ability of a test to differentiate between healthy and sick instances is a measure of its accuracy. Find the proportion of analysed cases with true positives and true negatives to get a sense of the test's accuracy. Based on the calculations:

Accuracy = TP + TN /(TP + TN + FP + FN)

$$Accuracy = \frac{(TN + TP)}{T}$$

Test Accuracy: 0.9895

Precision: The accuracy rate of a classification or number of positive cases is known as precision. Accuracy is determined by applying the following formula:

Precision = True positives/ (True positives + False positives) = TP/(TP + FP)

$$\Pr e\,cision = \frac{TP}{(TP + FP)}$$

Recall: The recall of a model is a measure of its capacity to identify all occurrences of a relevant machine learning class. A model's ability to detect class instances is shown by the ratio of correctly predicted positive observations to the total number of positives.

$$Recall = \frac{TP}{(FN + TP)}$$

mAP: One ranking quality statistic is Mean Average Precision (MAP). It takes into account the quantity of pertinent suggestions and where they are on the list. The arithmetic mean of the Average Precision (AP) at K for each user or query is used to compute MAP at K.

$$\boldsymbol{mAP = \frac{1}{n}\sum_{k=1}^{k=n} AP_k}$$

$$\boldsymbol{AP_k = \text{ the AP of class } k}$$

$$\boldsymbol{n = \text{ the number of classes}}$$

F1-Score: A high F1 score indicates that a machine learning model is accurate. Improving model accuracy by integrating recall and precision. How often a model gets a dataset prediction right is measured by the accuracy statistic..

$$F1 = 2 \cdot \frac{(Recall \cdot \Pr e\,cision)}{(Recall + \Pr e\,cision)}$$

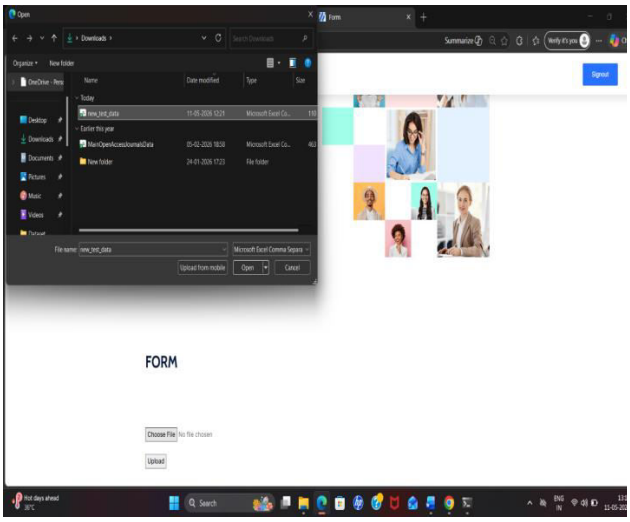

Fig 2:In above screen selecting and uploading test data and then click on 'upload' button to get below page

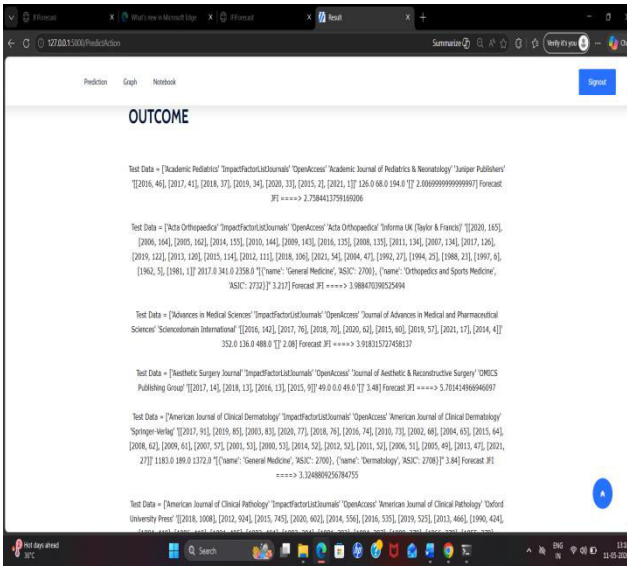

Fig 3:In above screen in square bracket can see Test data values and then after arrow =➔ can see predicted JFI (journal impact factor) values

5. CONCLUSION

This paper presented an effective machine learning approach for forecasting the Journal Impact Factor (JIF) in the presence of missing data. The study demonstrated that traditional imputation methods like mean substitution lead to lower prediction accuracy due to loss of data relationships. To overcome this limitation, a K-Nearest Neighbor Imputation (KNNI) technique was applied, which preserves the similarity between data points and improves data quality.

The experimental results confirmed that the optimized Support Vector Machine (SVM) model combined with KNNI achieved superior performance, producing lower RMSE and MAE compared to existing methods. This proves that proper handling of missing data along with model optimization plays a crucial role in improving prediction accuracy. The proposed system provides a reliable and efficient solution for JIF forecasting, helping researchers make better decisions in selecting appropriate journals for publication.

6. FUTURE SCOPE

The proposed system can be further enhanced by integrating advanced deep learning models such as Long Short-Term Memory (LSTM) and hybrid ensemble techniques to improve prediction accuracy. Incorporating additional journal evaluation metrics like h-index, SNIP, and CiteScore can provide a more comprehensive assessment of journal quality beyond JIF.

Future work can also explore more sophisticated missing data handling techniques, such as Iterative

Imputer and GAN-based imputation methods, to further improve data quality. Additionally, the system can be extended into a real-time web-based recommendation platform that suggests suitable journals for researchers based on their work. This would make the model more practical and impactful for real-world academic publishing applications.

REFERENCES

[1] J. Zhou, N. Cai, Z.-Y. Tan, and M. J. Khan, ‘‘Analysis of effects to journal impact factors based on citation networks generated via social computing,’’ IEEE Access, vol. 7, pp. 19775–19781, 2019.

[2] S. N. Groesser, ‘‘Dynamics of journal impact factors,’’ Syst. Res. Behav. Sci., vol. 29, no. 6, pp. 624–644, Nov. 2012. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/sres.2142

[3] S. T. H. Rizvi, M. Y. Latif, M. S. Amin, A. J. Telmoudi, and N. A. Shah, ‘‘Analysis of machine learning based imputation of missing data,’’ Cybern. Syst., pp. 1–15, 2023, doi: 10.1080/01969722.2023.2247257.

[4] C. Li, X. Ren, and G. Zhao, ‘‘Machine-learning-based imputation method for filling missing values in ground meteorological observation data,’’ Algorithms, vol. 16, no. 9, p. 422, Sep. 2023.

[5] Y. Zhang, B. Zhou, X. Cai, W. Guo, X. Ding, and X. Yuan, ‘‘Missing value imputation in multivariate time series with end-to-end generative adversarial networks,’’ Inf. Sci., vol. 551, pp. 67–82, Apr. 2021.

[6] J. M. B. Haslbeck, L. F. Bringmann, and L. J. Waldorp, ‘‘A tutorial on estimating time-varying vector autoregressive models,’’ Multivariate Behav. Res., vol. 56, no. 1, pp. 120–149, Jan. 2021.

[7] F. Khan, A. Saeed, and S. Ali, ‘‘Modelling and forecasting of new cases, deaths and recover cases of COVID-19 by using vector autoregressive model in Pakistan,’’ Chaos, Solitons Fractals, vol. 140, Nov. 2020, Art. no. 110189.

[8] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, ‘‘Informer: Beyond efficient transformer for long sequence time-series forecasting,’’ in Proc. AAAI Conf. Artif. Intell., vol. 35, May 2021, pp. 11106–11115.

[9] S. Du, T. Li, Y. Yang, and S.-J. Horng, ‘‘Multivariate time series forecasting via attention-based encoder–decoder framework,’’ Neurocomputing, vol. 388, pp. 269–279, May 2020.

[10] M. M. Öztürk, ‘‘Hyperparameter optimization of a parallelized LSTM for time series prediction,’’ Vietnam J. Comput. Sci., vol. 10, no. 3, pp. 303–328, Aug. 2023.

[11] X. Wen and W. Li, ‘‘Time series prediction based on LSTM-attention-LSTM model,’’ IEEE Access, vol. 11, pp. 48322–48331, 2023.

[12] F. Martínez, M. P. Frías, M. D. Pérez-Godoy, and A. J. Rivera, ‘‘Time series forecasting by generalized regression neural networks trained with multiple series,’’ IEEE Access, vol. 10, pp. 3275–3283, 2022.

[13] X. Song, Y. Liu, L. Xue, J. Wang, J. Zhang, J. Wang, L. Jiang, and Z. Cheng, ‘‘Time-series well performance prediction based on long short-term memory (LSTM) neural network model,’’ J. Petroleum Sci. Eng., vol. 186, Mar. 2020, Art. no. 106682.

[14] S. D. Yang, Z. A. Ali, H. Kwon, and B. M. Wong, ‘‘Predicting complex erosion profiles in steam distribution headers with convolutional and recurrent neural networks,’’ Ind. Eng. Chem. Res., vol. 61, no. 24, pp. 8520–8529, Jun. 2022.

[15] S. D. Yang, Z. A. Ali, and B. M. Wong, ‘‘FLUID-GPT (fast learning to understand and investigate dynamics with a generative pretrained transformer): Efficient predictions of particle trajectories and erosion,’’ Ind. Eng. Chem. Res., vol. 62, no. 37, pp. 15278–15289, Sep. 2023.

[16] A. Jadhav, D. Pramod, and K. Ramanathan, ‘‘Comparison of performance of data imputation methods for numeric dataset,’’ Appl. Artif. Intell., vol. 33, no. 10, pp. 913–933, Aug. 2019.

[17] G. Kabir, S. Tesfamariam, J. Hemsing, and R. Sadiq, ‘‘Handling incomplete and missing data in water network database using imputation methods,’’ Sustain. Resilient Infrastruct., vol. 5, no. 6, pp. 365–377, Nov. 2020.

[18] S. Daberdaku, E. Tavazzi, and B. Di Camillo, ‘‘A combined interpolation and weighted K-nearest neighbours approach for the imputation of longitudinal ICU laboratory data,’’ J. Healthcare Informat. Res., vol. 4, no. 2, pp. 174–188, Jun. 2020.

[19] J. C. Jakobsen, C. Gluud, J. Wetterslev, and P. Winkel, ‘‘When and how should multiple imputation be used for handling missing data in randomised clinical trials—A practical guide with flowcharts,’’ BMC Med. Res. Methodol., vol. 17, no. 1, pp. 1–10, Dec. 2017.

[20] R. Wu, S. D. Hamshaw, L. Yang, D. W. Kincaid, R. Etheridge, and A. Ghasemkhani, ‘‘Data imputation for multivariate time series sensor data with large gaps of missing data,’’ IEEE Sensors J., vol. 22, no. 11, pp. 10671–10683, Jun. 2022.

[21] Y. Luo, ‘‘Evaluating the state of the art in missing data imputation for clinical data,’’ Briefings Bioinf., vol. 23, no. 1, Jan. 2022, Art. no. bbab489.

[22] A. Zainuddin, M. A. Hairuddin, A. I. M. Yassin, Z. I. A. Latiff, and A. Azhar, ‘‘Time series data and recent imputation techniques for missing data: A review,’’ in Proc. Int. Conf. Green Energy, Comput. Sustain. Technol. (GECOST), Oct. 2022, pp. 346–350.

[23] R. Feng, D. Grana, and N. Balling, ‘‘Imputation of missing well log data by random forest and its uncertainty analysis,’’ Comput. Geosci., vol. 152, Jul. 2021, Art. no. 104763.

[24] Q.-T. Phan, Y.-K. Wu, Q.-D. Phan, and H.-Y. Lo, ‘‘A study on missing data imputation methods for improving hourly solar dataset,’’ in Proc. 8th Int. Conf. Appl. Syst. Innov. (ICASI), Apr. 2022, pp. 21–24.

[25] C. C. Nguyen, C. T. Tran, and B. N. Vi, ‘‘Deep learning for simultaneous imputation and classification of time series incomplete

data,'' J. Sci. Techn.-Sect. Inf. Commun. Technol., vol. 12, no. 1, pp. 110–125, 2023, doi: 10.56651/lqdtu.jst.v12.n1.661.ict.

[26] M. Sangeetha and M. S. Kumaran, ''Deep learning-based data imputation on time-variant data using recurrent neural network,'' Soft Comput., vol. 24, no. 17, pp. 13369–13380, Sep. 2020.

[27] A. J. Saroj, A. Guin, and M. Hunter, ''Deep LSTM recurrent neural networks for arterial traffic volume data imputation,'' J. Big Data Anal. Transp., vol. 3, no. 2, pp. 95–108, Aug. 2021.

[28] A. Flores, H. Tito-Chura, and V. Yana-Mamani, ''Wind speed time series imputation with a bidirectional gated recurrent unit (GRU) model,'' in Proc. Future Technol. Conf. (FTC), vol. 2. Vancouver, BC, Canada: Springer, 2022, pp. 445–458, doi: 10.1007/978-3-030-89880-9_34.

[29] A. Andreadis, ''Missing values imputation on multivariate time series in the field of agriculture,'' Fac. Sci., School Inform., Aristotle Univ. Thessaloniki, Thessaloniki, Greece, 2022.

[30] W. Du, ''A deep learning model to impute missing data in time series,'' Dept. Elect. Comput. Eng., Concordia Univ., Montreal, QC, Canada, 2021.

Author Profiles

Mrs.K.Pavani Working as Assistant & Head of Department of MCA ,in SRK Institute of technology in Vijayawada. She done with MCA ,M. Tech in Computer Science .She has 10 years of Teaching experience in SRK Institute of technology, Enikepadu, Vijayawada,NTR District. Her area of interest includes AI ML, etc

Ms.M.Lakshmi Durga is an MCA Student in the Department of Computer Application at SRK Institute Of Technology, Enikepadu, Vijayawada, NTR District. She has Completed Degree in B.Sc.(Mathematics,physics,computer science) from Sree Vidya Degree College Gudivada. Her area of interest are DBMS and Machine Learning with Python.