Grenze International Journal of Engineering and Technology, Jan Issue

Prediction in Multivariate Time Series Data using Generative Adversarial Network

Shrabani Medhi[1], Samudra Kashyap[2] and Nitul Das[3]
[1]Girijananda Chowdhury University/Department of Computer Science and Engineering, Guwahati, India
Email: shrabani_cse@gcuniversity.ac.in
[2-3]Girijananda Chowdhury Institute of Management and Technology/ Department of Computer Science and Engineering, Guwahati, India
Email: {samudrakashyap77@gmail.com, nituldas8753@gmail.com}

***Abstract*—Air pollution is an issue of great concern. PM2.5 is the most dangerous pollutant out of all the pollutants. A large number of missing values is present in multivariate pollution data. This makes the prediction of PM2.5 concentration very difficult. Traditional approaches to deal with missing values includes mean imputation, median imputation, case deletion, matrix factorization-based imputation, case deletion, etc. All these methods are not capable of modelling the temporal dependencies and complex distributed nature of multivariate pollution data. We have dealt with missing values by mean imputation, median imputation, Generative Adversarial Network imputation and Conditional Tabular Generative Adversarial Network. We have used Artificial Neural Network to predict PM2.5 concentration. The results reveal that for time series data, imputation by median model has performed waste. No significance improvement is seen in prediction of PM2.5 concentration using GAN and CTGAN model if we compare it with imputation by mean model. For time series data, more research work needs to be done with GAN and its variants so that better results can be achieved.**

***Index Terms*— PM2.5 prediction, Generative Adversarial Network, Conditional Tabular Generative Adversarial Network, data pre-processing.**

I. INTRODUCTION

Nowadays air pollution is a major global problem[1]. Pollution is divided into two types: natural pollution caused by volcanic eruptions and fires in forests causing emission of SO2, CO2, CO, NO2, and sulfates and man-made pollution caused by human activities such as burning of oils, discharges from industrial production processes, emissions from transportations that have PM2.5 as its major air pollutant. PM2.5 has destructive effects on human health, other kinds of creatures, and environment[2]. Air pollution leads to diseases related to cardiovascular and respiratory problems leading to death of animals and plants, acid rain, climate change, global warming, etc. That is why prediction of PM2.5 beforehand is very important.[3] However, prediction using multivariate pollution data is difficult due to problems in sensor devices, collection errors, etc leading to high amount of missing values[4]. The high range of missing values makes it very difficult to predict PM2.5 concentration. Therefore it is very important to deal with missing values in multivariate time series[5].
The missing values of the dataset can be handled in two ways. First way is to directly model the dataset with missing values [6]. Second way is to perform imputation of missing values to obtain the complete dataset and

Grenze ID: 01.GIJET.10.1.516

then use machine learning methods to perform prediction. Various conventional missing value imputation methods are available. In case deletion method [7], [8] observations with missing values are discarded. Other missing value imputation methods are median imputation, mean imputation, most common value imputation, etc. The main drawback of these statistical imputation methos is the incapability to utilize temporal data. Many machine learning based imputation methods [9]–[11] like maximum likelihood Expectation-Maximization (EM) based imputation, K Nearest Neighbour (KNN) based imputation and Matrix Factorization based imputation. The drawback of these methods is that they are incapable to deal with the temporal relations between observations.

Generative Adversarial Networks (GAN) was introduced in [12] where the latent distribution of dataset is learnt and real samples are generated from random noise. GAN is successfully used in many applications like face completion, sentence generation, etc[13]–[15]. All these applications are based on non-sequential dataset and very less work is adopted on datasets related to temporal relations [16]. GAN has achieved lots of success in image imputation. By taking inspiration from this we try to generate and impute original incomplete multivariate air pollution data. We impute the missing values using mean, median, GAN and CTGAN. Then we apply Artificial Neural Network (ANN) to predict PM2.5 prediction.

II. Literature Review

Related studies done is shown in Table I.

III. Proposed Methodology

The workflow diagram of our paper is given in Fig.1.

A. Data Source

The dataset for our study is collected from used Pollution Control Board, Assam located in Bamunimaidam which gives the Continuous Ambient Air Quality Monitoring Station (CAAQMS) data for Guwahati city. The data set contains hourly data of 2 years. The parameters used in the study are given in Table.2. Meteorological conditions, criteria gases and particulates are used as parameters. Total 9 features are used. Table 4 provides a descriptive statistic of the available meteorological conditions, criteria gases and particulates measures: count, mean, standard deviation, min, 25%, 50%, 75%, maximum, skewness, kurtosis and variance. There is no high value of skewness in data. It indicates that there is no sharp increase in the data. The high value of kurtosis in PM2.5 indicates the presence of data discontinuities.

B. Data Preprocessing

Data pre-processing is a crucial step in preparing raw data to make it suitable for machine learning models. In our study, we performed several pre-processing steps to ensure the quality and compatibility of the dataset for our machine learning tasks[17]. Two most important issues that we need to deal with are missing values and outliers[18]. We have developed 4 models by following four procedures for dealing with missing values: imputation by mean (IMmean), imputation by median (IMmedian), imputation by GAN(IMGAN), imputation by CTGAN (IMCTGAN). We have used StandardScalar for normalization.

C. Imputation by GAN

Generative Adversarial Networks (GANs) are a type of deep learning architecture that consists of two main components: a generator and a discriminator. The goal of GANs is to learn the underlying probability distribution of a given dataset and generate new samples that are similar to the training data[19].

In the context of regression-type datasets, GANs can be used to generate synthetic data points that follow a similar distribution to the original dataset[20]. This can be particularly useful when the available dataset is limited or when collecting new data is expensive or time-consuming. The workflow of Imputation by GAN model is shown in Fig.2.

D. Imputation by CTGAN

Conditional Tabular Generative Adversarial Networks (CTGAN) [21] are an extension of Generative Adversarial Networks (GANs) that are specifically designed for generating synthetic tabular data. This technique is particularly useful when dealing with regression-type datasets, where the goal is to predict a continuous target variable based on a set of input features[22].

TABLE I. LITERATURE REVIEW

SL No.	Algorithm applied	Parameters used	Other techniques applied	Evaluation metrics used	Result
1	Membership inference attacks, model inversion attacks, model poisoning attacks, backdoor attacks[23]	Attack type, data and model characteristics, attack success metrics, defense, mechanisms, evaluation metrics,	Differential privacy, secure multi-party computation (smc), federated learning, adversarial training, privacy-preserving data perturbation	Attack success rate, accuracy, false positive rate and false negative rate, area under the curve (AUC)	Proposed model is applied to 1-nn dtw and fully connected network (FCN). It is trained on 42 university of california riverside (UCR) datasets.
2	Query the Black-Box Model, Adversarial Example Generation, Adversarial Example Evaluation [24]	Perturbation magnitude, Step size, Number of iterations	One-hot encoding, Targeted loss function, Class balancing, Transferability	Success Rate, Confidence Level, Fooling Rate, Robustness Evaluation, Transferability	Adversarial perturbations can reduce the accuracy of deep learning classifier (RESNET).
3	Smooth Gradient Method (SGM), Gradient Method (GM), way of countering SGM attacks.[25]	Gaussian filter parameters, Perturbation magnitude, Adversarial attack target, Model architecture	Iterative FGSM, Projected Gradient Descent (PGD), Waveform Transformation Attack, Differentiable Augmentation	Success rate, Attack strength, Perturbation size, Imperceptibility, Robustness	Results reveal that attacks are more likely to be detected on time series. There is need smoother perturbations.
4	Adversarial training, Defensive distillation, Gradient masking, Randomization, Feature squeezing, Ensemble [26]	Perturbation size, Attack budget, Attack type, Input space, Attack objective, Attack surface	Model diversification, Input preprocessing, Model introspection, Adversarial retraining, Ensemble defenses, Game-theoretic defenses	Success rate, Attack success rate, False positive rate, Robust accuracy, Detection rate, Mean perturbation, Confidence score	The overall results of evasion attacks on machine learning models during testing time can be of concern and focuses the need for more robust defenses against such attacks.
5	Data Preparation, Training the GAN, Imputing Missing Data, Evaluation[27]	Learning rate, Number of epochs, Number of epochs, Network architecture, Loss function, Input noise, Dropout rate	Preprocessing techniques, Hybrid approaches, Transfer learning, post-processing techniques, Ensemble techniques	Mean squared error (MSE), Root mean squared error (RMSE), Mean absolute error (MAE), Coefficient of determination (R^2), Pearson correlation coefficient, Structural similarity index (SSIM)	GAN-based imputation can be effective for filling up missing values in multivariate time series data. The performance of the model depends on factors such as the specific dataset, quantity and pattern of missing data, and hyperparameters used.

CTGAN is designed to generate synthetic data that not only captures the statistical properties of the original dataset but also preserves the conditional relationships between input features and the target variable. In other words, CTGAN can generate new data points with similar feature distributions and target variable distributions as the original dataset. The workflow diagram of Imputation by CTGAN model is shown in Fig.3.

TABLE II. SUMMARY OF MEASUREMENT SITE AND OBSERVED VARIABLES

Measurement site	Type	Variables
Guwahati city	Meteorological conditions	Relative humidity, Wind speed, Wind direction, Temperature, Rainfall
	Criteria gases	NO_2, SO_2,
	Particulates	PM2.5, PM10

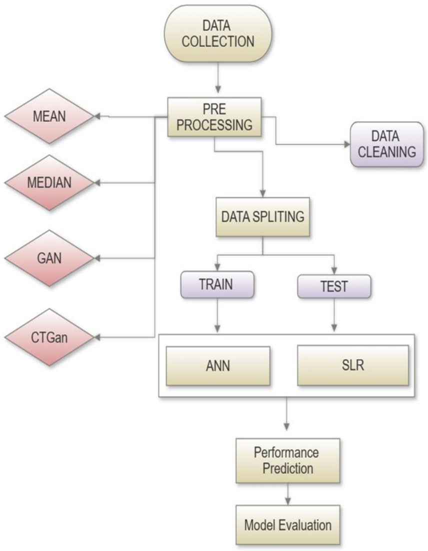

Figure 1. Workflow diagram

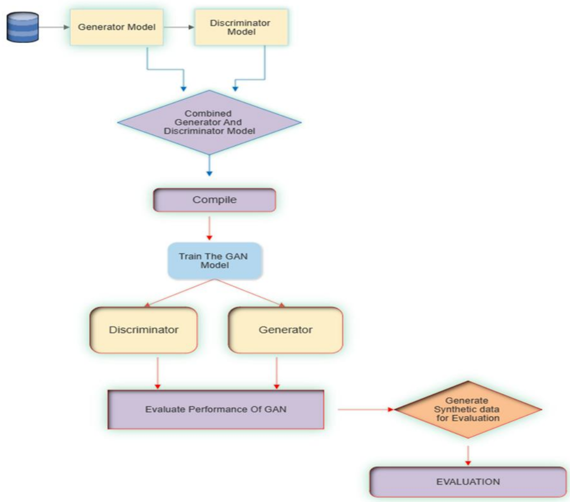

Figure 2. IMGAN Model

E. Model Training

Scikit-Learn library and Python library was used to create 4 models using imputation by mean, imputation by median, GAN and CTGAN. We have used ANN[28]**.** After defining the models and compiling them with appropriate loss functions and optimizers, we proceed to train the GAN model. For the Generator Model ReLU activation is used and tanh activation function is used in output layer. For Discrimantor Model leaky ReLU is used in dense layers and sigmoid activation function is used in final output layer. For the Combined Generator and Discriminator Model, binary cross-entropy loss and Adam optimizer is used. We have also applied hyperparameter tuning using GridSearchCV. The grid search was conducted with cross-validation using a 3-fold split. We have split the data into training and testing sets with a ratio of 80:20. To further enhance the performance of our model, we employed early stopping as a form of regularization. Early stopping allows us to stop training if the validation loss does not improve over a certain number of epochs. We defined an EarlyStopping callback and incorporated it into the model fitting process. The model was trained for a maximum of 100 epochs with a batch size of 32. We also included a validation split of 0.2 to assess the model's performance on unseen data during training. We have trained CTGAN model on the original dataset to learn the underlying data distribution and then generating synthetic data that closely resembles the original dataset. The synthetic data is evaluated using linear regression models and compared with the original dataset to assess its utility and effectiveness. The CTGAN model is trained on the original dataset to learn the underlying data distribution. This step allows the CTGAN model to capture the patterns and dependencies present in the real data, which it will use to generate synthetic data that resembles the original data.

IV. Results and Discussion

Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), root mean squared logarithmic error (RMSLE) and R-squared (R2) score are computed as metrics to evaluate the four models created by using imputation by mean, imputation by median, GAN and CTGAN.

The R2 score provides an indication of how well the model fits the data, with a higher value indicating a better fit. The MAE measures the average absolute difference between the predicted and actual values, while the MSE calculates the average squared difference. RMSE, the square root of MSE, is commonly used as a measure of the model's prediction error. RMSLE is particularly useful when the target variable has a wide range of values. The R2 score, MAE, MSE, RMSE and RMSLE for all the four models, namely, imputation by mean, imputation by median, GAN and CTGAN is shown in Fig. 4, 5, 6, 7, respectively. The prediction accuracy score of all the four models, namely, imputation by mean, imputation by median, GAN and CTGAN is shown in Fig.8. Finally, in Fig.9, the accuracy score of all the models is compared based on R2 score, MAE, MSE, RMSE and RMSLE. We observe that the accuracy score of models, imputation by mean, GAN and CTGAN is almost same. The accuracy score of imputation by median is low compared to the other models. The MAE score of imputation by median model is highest and lowest is achieved by model, imputation by mean. MAE of model by GAN and CTGAN

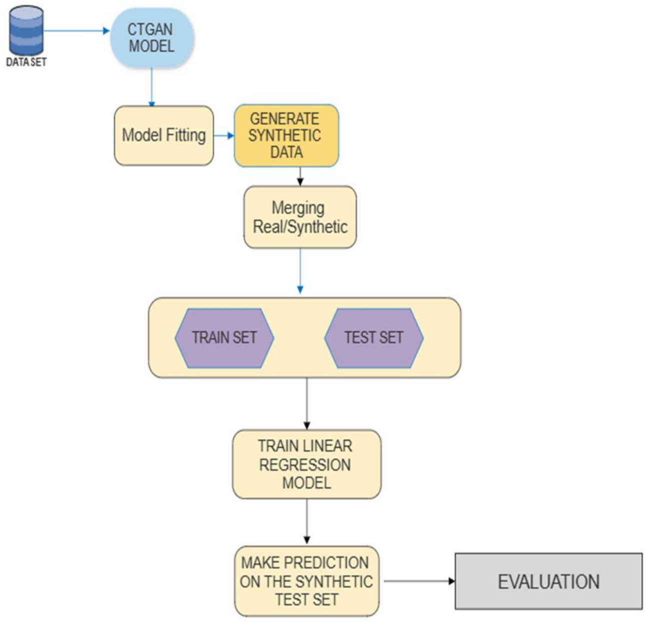

Figure 3. Imputation by IMCTGAN Model

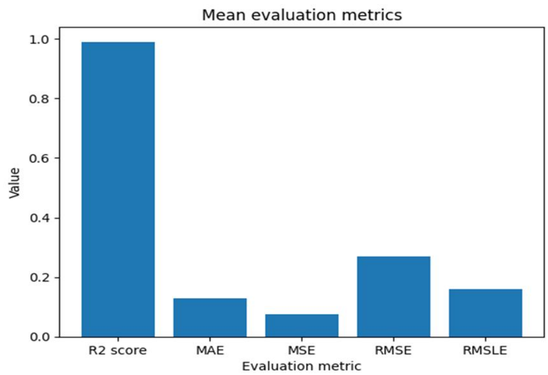

Figure 4. Evaluation metrics for IMmean model

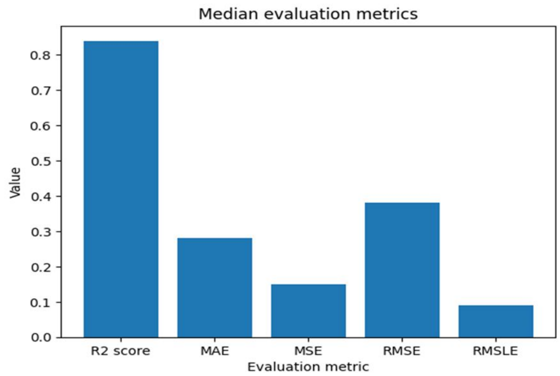

Figure 5 Evaluation metrics for IMmedian model

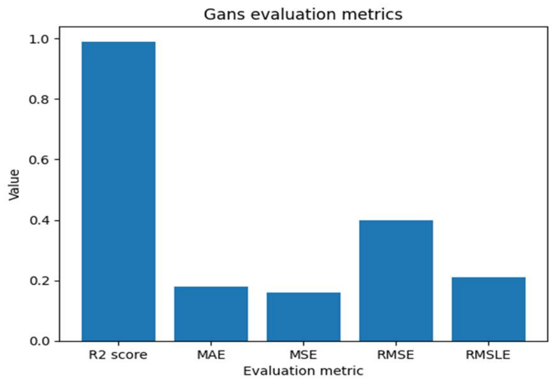

Figure 6. Evaluation metrics for IMGAN model

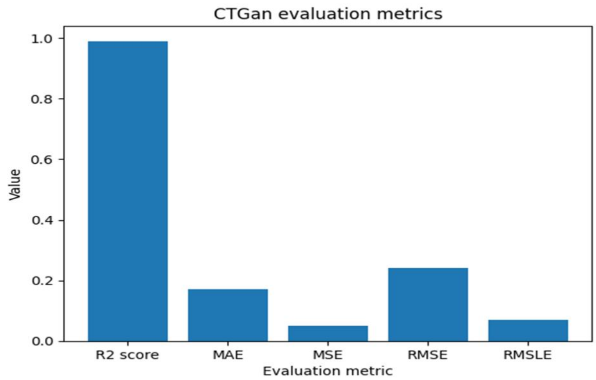

Figure 7 Evaluation metrics for IMCTGAN model

has performed average with CTGAN slightly better than GAN. However, MSE value is highest for model with GAN and lowest for CTGAN. On the other hand, the RMSE value of GAN and imputation by median model is way higher than imputation by mean and CTGAN model. The comparison of all the models is shown in Table III.

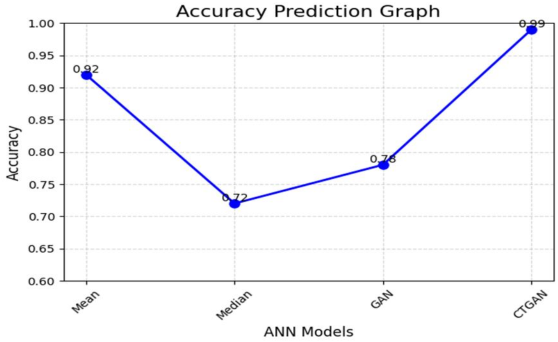

Fig.8. Prediction Accuracy Score

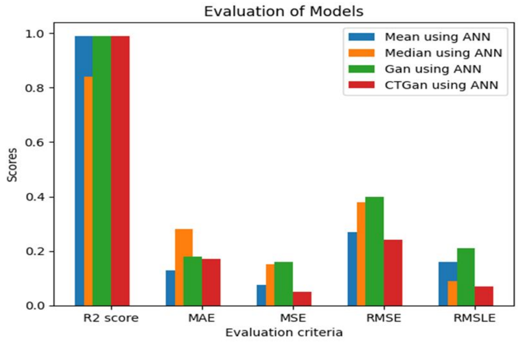

Fig.9. Evaluation Criteria of the Models

TABLE III. EVALUATION METRICS COMPARISON

Metrics	IMmean	IMmedian	IMGAN	IMCTGAN
R2 score	0.990	0.84	0.99	0.99
MAE	0.130	0.28	0.18	0.17
MSE	0.074	0.15	0.16	0.05
RMSE	0.270	0.38	0.40	0.24
RMSLE	0.160	0.09	0.21	0.07

V. CONCLUSION

From the results we can conclude that for time series data, imputation by median model has performed worst. Performance of Imputation by Mean model is best among all the models. Imputation by GAN and Imputation by CTGAN model has performed average. No significance improvement is seen in prediction of PM2.5 concentration using GAN and CTGAN model if we compare it with imputation by mean model. For time series data, more research work needs to be done with GAN and its variants so that better results can be achieved. More research in this area needs to be done with the variants of GAN for the prediction of PM2.5 concentration in multivariate time series data.

ACKNOWLEDGMENT

First and foremost, we would like thank almighty GOD for everything that happens to us and makes us patient, dedicated and courageous to our work. We are very grateful to Pollution Control Board, Assam located in Bamunimaidan for providing us the necessary dataset without which it was not possible for us to carry out our research. Finally, we wish to warmly thank our institution “Girijananda Chowdhury Institute of Management and Technology-Guwahati” for encouraging us to do research work in our field.

REFERENCES

[1] X. Li, L. Jin, and H. Kan, “Air pollution: a global problem needs local fixes,” Nature, vol. 570, no. 7762, pp. 437–439, 2019.

[2] Y.-F. Xing, Y.-H. Xu, M.-H. Shi, and Y.-X. Lian, “The impact of PM2. 5 on the human respiratory system,” Journal of thoracic disease, vol. 8, no. 1, p. E69, 2016.

[3] S. Medhi and M. Gogoi, “Visualization and Analysis of COVID-19 Impact on PM2. 5 Concentration in Guwahati city,” in 2021 International Conference on Computational Performance Evaluation (ComPE), IEEE, 2021, pp. 012–016.

[4] P. J. García-Laencina, J.-L. Sancho-Gómez, and A. R. Figueiras-Vidal, “Pattern classification with missing data: a review,” Neural Computing and Applications, vol. 19, pp. 263–282, 2010.

[5] P. J. García-Laencina, P. H. Abreu, M. H. Abreu, and N. Afonoso, “Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values,” Computers in biology and medicine, vol. 59, pp. 125–133, 2015.

[6] K. Zheng, J. Gao, K. Y. Ngiam, B. C. Ooi, and W. L. J. Yip, “Resolving the bias in electronic medical records,” in Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017, pp. 2171–2180.
[7] L. O. Silva and L. E. Zárate, “A brief review of the main approaches for treatment of missing data,” Intelligent Data Analysis, vol. 18, no. 6, pp. 1177–1198, 2014.
[8] J. Kaiser, “Dealing with Missing Values in Data.,” Journal of Systems Integration (1804-2724), vol. 5, no. 1, 2014.
[9] G. E. Batista and M. C. Monard, “An analysis of four missing data treatment methods for supervised learning,” Applied artificial intelligence, vol. 17, no. 5–6, pp. 519–533, 2003.
[10] T. Hastie, R. Mazumder, J. D. Lee, and R. Zadeh, “Matrix completion and low-rank SVD via fast alternating least squares,” The Journal of Machine Learning Research, vol. 16, no. 1, pp. 3367–3402, 2015.
[11] R. Mazumder, T. Hastie, and R. Tibshirani, “Spectral regularization algorithms for learning large incomplete matrices,” The Journal of Machine Learning Research, vol. 11, pp. 2287–2322, 2010.
[12] I. Goodfellow et al., “Generative adversarial networks,” Communications of the ACM, vol. 63, no. 11, pp. 139–144, 2020.
[13] A. Bora, E. Price, and A. G. Dimakis, “AmbientGAN: Generative models from lossy measurements,” in International conference on learning representations, 2018.
[14] Y. Li, S. Liu, J. Yang, and M.-H. Yang, “Generative face completion,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 3911–3919.
[15] Z. Lu, Z. Li, J. Cao, R. He, and Z. Sun, “Recent progress of face image synthesis,” in 2017 4th IAPR Asian Conference on Pattern Recognition (ACPR), IEEE, 2017, pp. 7–12.
[16] J. Yoon, J. Jordon, and M. Schaar, “Gain: Missing data imputation using generative adversarial nets,” in International conference on machine learning, PMLR, 2018, pp. 5689–5698.
[17] C. V. G. Zelaya, “Towards explaining the effects of data preprocessing on machine learning,” in 2019 IEEE 35th international conference on data engineering (ICDE), IEEE, 2019, pp. 2086–2090.
[18] J. Burdack, F. Horst, S. Giesselbach, I. Hassan, S. Daffner, and W. I. Schöllhorn, “Systematic comparison of the influence of different data preprocessing methods on the performance of gait classifications using machine learning,” Frontiers in bioengineering and biotechnology, vol. 8, p. 260, 2020.
[19] P. Yazdanian and S. Sharifian, “E2LG: a multiscale ensemble of LSTM/GAN deep learning architecture for multistep-ahead cloud workload prediction,” The Journal of Supercomputing, vol. 77, pp. 11052–11082, 2021.
[20] A. E. Bakhshipour, A. Koochali, U. Dittmer, A. Haghighi, S. Ahmad, and A. Dengel, “A Bayesian Generative Adversarial Network (GAN) to Generate Synthetic Time-Series Data, Application in Combined Sewer Flow Prediction,” arXiv preprint arXiv:2301.13733, 2023.
[21] F. Li, H. Zheng, and X. Li, “A novel hybrid model for multi-step ahead photovoltaic power prediction based on conditional time series generative adversarial networks,” Renewable Energy, vol. 199, pp. 560–586, 2022.
[22] K. E. Smith and A. O. Smith, “Conditional GAN for timeseries generation,” arXiv preprint arXiv:2006.16477, 2020.