Electricity Load Forecasting using Hybrid Datasets with Linear Interpolation and Synthetic Data

Karma Dorji
School of Renewable Energy and Smart Grid Technology (SGtech), Naresuan University, Phitsanulok, Thailand
karmad65@nu.ac.th

Sorawut Jittanon
School of Renewable Energy and Smart Grid Technology (SGtech), Naresuan University, Phitsanulok, Thailand
sorawutj66@nu.ac.th

Prapita Thanarak
School of Renewable Energy and Smart Grid Technology (SGtech), Naresuan University, Phitsanulok, Thailand
prapitat@nu.ac.th

Pornthip Mensin
School of Renewable Energy and Smart Grid Technology (SGtech), Naresuan University, Phitsanulok, Thailand
pornthipw@nu.ac.th

Chakkrit Termritthikun
School of Renewable Energy and Smart Grid Technology (SGtech), Naresuan University, Phitsanulok, Thailand
chakkritt@nu.ac.th (corresponding author)

Received: 31 July 2024 | Revised: 23 September 2024 | Accepted: 28 September 2024

ABSTRACT

Electricity load forecasting is an important aspect of power system management. Improving forecasting accuracy ensures reliable electricity supply, grid operations, and cost savings. Often, collected data consist of Missing Values (MVs), anomalies, outliers, or other inconsistencies caused by power failures, metering errors, data collection errors, hardware failures, network failures, or other unexpected events. This study uses real-world data to investigate the possibility of using synthetically generated data as an alternative to filling in MVs. Three datasets were created from an original one based on different imputation methods. The imputation methods employed were linear interpolation, imputation using synthetic data, and a proposed hybrid method based on linear interpolation and synthetic data. The performance of the three datasets was compared using deep learning, machine learning, and statistical models and verified based on forecasting accuracy improvements. The findings demonstrate that the hybrid dataset outperformed the other interpolation methods based on the forecasting accuracy of the models.

Keywords-bad data; missing values; deep learning; synthetic data; electricity load forecasting; generative adversarial network

I. INTRODUCTION

Managing electricity demand and supply is a crucial aspect of a power system. Every country relies on energy to develop its industries and economy [1]. Energy is a real-time resource with limited storage capacity. Forecast errors lead to wasted resources and increased operational costs. Thus, the quest for improved accurate forecasting models is an ongoing process. As a single generic model cannot effectively address all issues, the forecasting problem is categorized into short-, medium-, and long-term load forecasting [2]. Machine Learning (ML) and Deep Learning (DL) based load predictions have experienced explosive growth in recent years due to their ability to handle nonlinearity, large data, feature extraction automation, and good performance [3].

Energy management systems use smart meters to collect fine-grained consumption readings from houses, buildings, towns, and cities and perform operations such as load and demand forecasting [4]. Load data are a key component in load forecasting, which must be cleaned by removing errors and handling missing data to train and test forecasting models [5]. The collected data often have problems, such as missing data, outliers, anomalies, and other inconsistencies. The main causes of these problems are power failures, measurement errors, data collection errors, hardware failures, network failures, or other unexpected events [6]. In any real-world data, missing data is a ubiquitous problem. Regardless of the DL or ML model employed, handling missing data is an important issue wherever data quality cannot be ensured. Models trained on limited or low-quality data decrease the model's accuracy [7, 8]. Therefore, it is essential to impute missing data and process them to obtain better and more accurate forecasts [8, 9]. Missing data are commonly classified into three categories. The first is Missing Completely At Random (MCAR), which denotes entirely random missing data points that do not depend on observed or unobserved values. Missing At Random (MAR) describes missing data related to the observed values, but not to the missing ones. Missing Not At Random (MNAR) describes missing data that depend on observed and unobserved values [10]. In this study, the missing data were considered MCAR and MNAR.

Missing data problems have been studied across various sectors. Imputation methods are commonly classified into three types. The first is case deletion, which ignores some important information and discards incomplete observations. This approach results in poor results with an increasing missing rate. The second is statistical imputation, which employs statistical algorithms, such as mean, median, or the most common value imputation. The third approach uses ML and DL algorithms to make predictions to impute MVs, also called predictive methods [4, 11]. Despite advances in missing data imputation using statistical, ML, and DL techniques, challenges remain. Some studies presented imputation models focusing on specific datasets, as each dataset has unique characteristics. The performance of various methods differs with varying missing rates [12], which requires a study to handle different missing data characteristics using specific techniques. Missing data are single missing points and missing block types. The type of single missing points are isolated single missing points with one record (hour) missing. On the other hand, missing blocks consist of consecutive missing records, as shown in Figure 1.

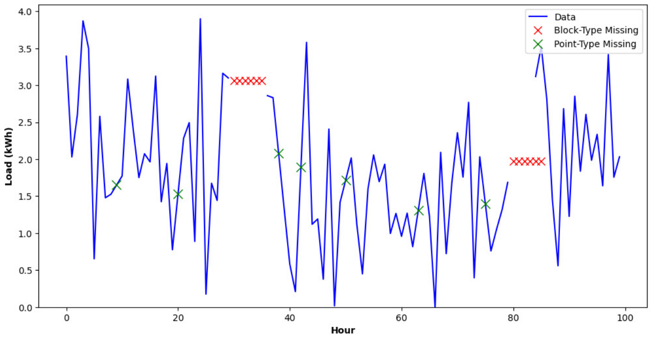

Fig. 1. Types of missing blocks. A green × represents a single missing point type, while consecutive red ××× represents missing blocks.

Most studies on missing data imputation generate the missing data artificially [12-15], and the developed model is used to impute or reconstruct the entire missing data. Studies such as [12, 16] stated that linear interpolation is effective when missing gaps are smaller, but performance deteriorates when missing gaps and missing rates increase. This study focuses mainly on imputing the smaller missing gaps with linear interpolation and the larger missing blocks with Generative Adversarial Networks (GANs). GANs are one of the approaches used for missing data imputation. GANs are typically used as a tool for data augmentation to reconstruct missing or limited data.

Multiple methods have been proposed to generate synthetic data, with the most popular being Variational Auto-Encoders (VAEs) and GANs. This study utilized the TimeGAN model, proposed in [17]. TimeGAN was selected because it can handle sequential time series data and is publicly available. The dataset obtained from these imputation methods was used to train the models and compare their forecasting accuracy to determine which imputation method is the best for this dataset. The dataset consisted of two years of data from SGtech's prosumer building. The contributions of this work are the following:

- A new hybrid approach to impute missing data, based on linear interpolation and synthetic data, explores the applicability of GANs in imputing missing data in electricity load forecasting. Handle effectively different missing data, considering the missing points and blocks of a dataset.
- The proposed hybrid method outperforms the linear interpolation and synthetic data missing data imputation methods in terms of forecasting accuracy improvement of the implemented models.

II. RELATED WORKS

This section provides an overview of the two key components of this study: Linear Interpolation (LI) and GANs.

A. Linear Interpolation (LI)

LI estimates a missing value m_i from the closest preceding and succeeding available values m_h and m_j . The generic representation of LI is

$$m_i = m_h + \frac{(x-x_i)*(m_j-m_h)}{(x_j-x_i)} \tag{1}$$

where $x_i < x < x_j$. x_i and x_j are the indices of the known data points surrounding the missing data point, $m_j - m_h$ are the values of the known data points, x is the position of missing data and m_i is the missing data that will be imputed.

LI estimates the value at other points through all known points. Implementing LI is a simple and fast process [12, 18, 19]. As the missing rate increases, the performance of LI imputation decreases. However, this is true for all imputation methods. According to [12], LI and the proposed CC-GAIN model exhibited the best performance compared to the other imputation methods. In [20], three interpolation methods were compared, namely linear, quadratic, and cubic interpolation, on a PM_{10} dataset, where the linear interpolation method achieved the best results. Similarly, in [14], it was found that linear interpolation was more accurate for short-term solar irradiance forecasting. Therefore, this study chose LI as the baseline imputation model to compare with the synthetic data imputation method. Moreover, since LI shows strong performance in lower missing data points, LI was used to impute point-type missing data, and synthetic data were used to impute the larger block missing points when creating the proposed HF dataset.

B. Generative Adversarial Networks (GANs)

GANs were initially developed for image generation. Their success has inspired fields such as natural language processing for tasks such as sentence generation. GANs have also been used for time series problems, such as imputing missing data and data augmentation in medical research, electricity load, solar data, and traffic data. The GAN framework typically consists of two neural networks, a generator G and a discriminator D. The generator network generates artificial data from the training data, and the discriminator network tries to differentiate if the generated data are fake or real [21].

Several studies have proposed specific GAN models for missing data imputation in electricity loads. CC-GAIN, a missing data imputation model that combines unsupervised clustering and classification-based GAN [12], improved the accuracy of the imputation by maintaining the data characteristics effective for various missing data rates, outperforming other imputation methods. In [22], Least Squares Generative adversarial Networks (LSGAN) were proposed, which is an unsupervised learning model. In this approach, the network automatically learns the measurement data, the power distribution patterns, and other correlations, enabling the generator to generate highly accurate data to reconstruct the missing ones. The accuracy of the model was higher than that of the other GAN models it was compared with. In [23], an Augmented Neural Ordinary Differential Equation-assisted Generative Adversarial Network (ANODE-GAN) was proposed, which used VAE to map incomplete time series to latent vectors, generate continuous-time dynamics, and decode them into complete data. With an additional discriminative network, ANODE-GAN accurately imputed missing data while preserving original features and temporal dynamics.

However, most GAN implementations are closed-source, making it difficult to replicate and reproduce them. Reproduction of such works requires expertise in the field, which is a time-consuming and costly undertaking. This study implements a publicly available GAN-based model called TimeGAN [17] to generate synthetic time series to impute missing data in a dataset. TimeGAN is a supervised GAN model designed to model the temporal dependencies in sequential data, tailored to create realistic time-series data. TimeGAN has been used effectively for data augmentation in load forecasting problems. For instance, in [24], synthetic and real data were investigated for electricity load forecasting. TimeGAN was used to generate realistic artificial electricity load data, and the dataset was tested in three configurations based on a Gated Recurrent Unit (GRU) neural network. Experiments were conducted with synthetic, real, and a mixture of real and synthetic data. The results showed that TimeGAN is a viable option for data augmentation. TimeGAN has also been used for data augmentation in other time series problems, such as health, telecommunications, and stocks [25-27].

III. METHOD

A. Data Collection And Preprocessing

The data for this study were collected from the SGtech's prosumer building at hourly intervals. The dataset consisted of two years of records, from 1 January 2022 to 31 December 2023. The dataset contains two features: datetime and load. Additional temporal features, such as hour, hour_group, weekday, weekend, worktime, and holiday, were extracted from datetime. Outliers were removed according to the 3-sigma rule. The dataset consists of 14,388 actual observations, with 3,131 observations missing, as the total expected record should be 17,519 hours. The total percentage of missing data in the dataset was calculated by

$$BadData(\%) = \left(100 - \frac{AD}{ED}\right) \times 100 \tag{2}$$

where AD denotes the actual data and ED denotes the expected data. This totals 17.87% of missing values in the dataset.

B. Analysis of Missing Data in the Dataset

In [12, 28, 29], missing data were artificially generated by creating missing gaps of 10% to 90% of the dataset. In this study, missing data were due to the failure of the metering device to record data, and a large portion of them was due to a malfunction of the metering device caused by a lightning strike, which caused the meter to not record data for 656 continuous hours, making up 21% of the total missing data. Missing data are point- or block-type (clusters) [12]. In this dataset, there were 891 hours of point-type missing data spread across the dataset, which constitutes 28.5% of the total MVs. The rest of the missing data were block-type, ranging from 2 to a maximum of 656 continuous hours of missing observations between April and May 2022. Two missing blocks had more than 200 continuous missing data points, and two other blocks had more than 100 continuous missing data points. The missing blocks with more than 100 continuous missing points constituted about 42% of the entire MV. The rest of the missing data were blocks of more than 2 and less than 100, constituting approximately 30% of the MVs.

C. Synthetic Data Generation Process

The 2023 dataset was used as a sample to generate synthetic data since it contained the least MVs. GAN requires high-quality training data for stability and performance. It is important to have minimal missing data in the training dataset for a GAN. The reason is to introduce minimal bias and generate quality synthetic data [21, 30]. For the 2023 dataset, MVs were initially filled using the interpolation method, and the synthetic data were generated using the TimeGAN model. The total synthetic data generated was 17,519 hours, which is equivalent to the expected data from the original dataset. The synthetically generated data were sequential and held the datetime record of the original dataset. A statistical similarity test on the load feature of the synthetically generated data was performed with the real data. It consisted of five metrics: mean, standard deviation, median, 25% quantile, and 75% quantile. The similarities were between 0 and 1, with 1 representing equal values. For the load feature of the synthetically generated data, the statistical measures were 1.00 for all metrics, indicating perfect similarity between the real and the synthetic data.

D. Missing Data Imputation

The 2022 and 2023 datasets had 3,131 MV records. Missing data were imputed using three approaches. In the first approach, the LI method was used to fill all MVs. This dataset is called LI and is considered the baseline for this study. In the second approach, synthetically generated data were used to replace the MVs. This dataset is denoted as SF. In the third approach, MVs were filled using a mix or a hybrid of LI and SF. This dataset was denoted as HF. All MVs were replaced corresponding to the actual datetime sequence in the case of replacing missing data with synthetically generated data.

1) Missing Data Imputation Using Linear Interpolation (LI)

For the first dataset (LI), MVs were replaced using the linear interpolation method. LI results in a long sloping linear line when MVs are continuous, as the LI method calculates MVs by drawing a straight line between two known data points. LI is an accurate missing data imputation method when the missing blocks are smaller, and its accuracy decreases as the size of the missing block increases.

2) Missing Data Imputation Using Synthetic Data (SF)

For the second dataset (SF), MVs were filled using synthetically generated data using TimeGAN. All MVs were replaced corresponding to the datetime of the original and the synthetic data. The problem with this dataset is that when the missing points are low, the replaced MVs in most cases are very different from the values of their neighboring values. In reality, the replaced values should be closer to the previous and the next data points, since the electricity load does not change drastically from one observation to the next.

3) The Proposed Hybrid Missing Data Imputation by Linear Interpolation and Synthetic Data (HF)

The proposed data imputation method was developed due to the problems found with the above two approaches in imputing MVs. In the third dataset (HF), MVs were filled by combining LI and synthetic data. The reason for this was that it is known that LI gives better results when block MVs are short. Certainly, the electricity load does not change drastically from one hour to the next. When observing the missing data filled using only synthetic data, it can be seen that the filled MVs for some smaller blocks significantly differ from the original series.

This study experimented with different configurations of LI and synthetic data. The main goal was to identify the optimal combination of LI and SF to impute MVs, as illustrated in Figure 2. This process was iterated five times. In the first iteration, all point-type MVs were replaced with LI, while missing blocks of two or more consecutive values were imputed using SF. The dataset was then fed into the forecasting models, and the results were compared with the Mean Absolute Error (MAE) scores of the LI-imputed dataset. If the compared result is not satisfactory, the approach was modified in the next iteration, applying LI to missing blocks of two or fewer MVs while using SF for the remaining missing blocks. After the fifth iteration, it was determined that replacing blocks of three or fewer MVs with LI and blocks of four or more with SF achieved the best forecasting results. In the proposed HF dataset, 1,127 hours of data points with missing blocks of three or less MVs were replaced using the LI method, and the remaining 2,004 MVs were replaced using the SF method.

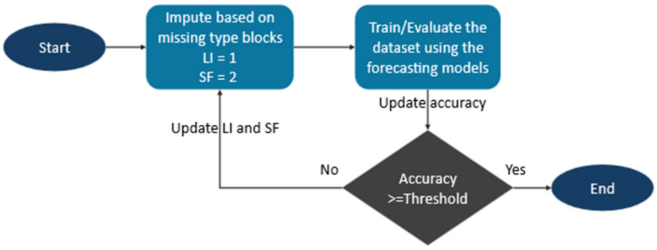
Start
Impute based on missing type blocks
LI = 1
SF = 2
Train/Evaluate the dataset using the forecasting models
Update accuracy
Update LI and SF
No
Accuracy >=Threshold
Yes
End

Fig. 2. Hybrid missing data imputation process.

E. Forecasting Models

The main intention was to verify if there are improvements using the imputed datasets with statistical, ML, and DL models. Auto-Regressive Integrated Moving Average (ARIMA), Support Vector Regression, and the Time-series Dense Encoder (TiDE) and Temporal Fusion Transformer (TFT) DL models were used. Each dataset was trained and tested using 80-20 train-test splits. First, the models were trained and tested on the LI dataset. The forecast accuracy result obtained was used as a baseline to compare the performance on the other two datasets. The performance of a dataset is considered good if the models have a better forecasting accuracy than the baseline. The forecast was carried out for multiple time horizons, ranging from 1, 2, 3, 5, and 7-day forecasts, to study the model's performance on different forecasting horizons.

1) Auto-Regressive Integrated Moving Average (ARIMA)

ARIMA is one of the most popular statistical approaches applied to electricity load forecasting. In general, ARIMA is written with the notation $ARIMA\ (p, d, q)$, where p denotes the autoregressive orders in the model, d is the order of differencing, and q is the moving average component. ARIMA also considers multiple seasonality, which is useful for electricity load forecasting.

2) Support Vector Regression (SVR)

SVR is a supervised ML method based on the Support Vector Machine (SVM) to model seasonal and cyclical effects. SVR, similar to neural network models, is capable of forming complex decision boundaries, but unlike them, SVR does not overfit the training data. It is effective for high-dimensional, nonlinear data.

3) Time-Series Dense Encoder (TiDE)

TiDE is a novel long-term time series DL forecasting model. TiDE uses a fully connected dense encoder that processes time series data efficiently, enabling it to capture long-term dependencies while being computationally efficient. TiDE can learn global as well as local patterns in time series data and ensures robust performance across different time horizons. TiDE is efficient in long-term forecast horizons.

4) Temporal Fusion Transformer (TFT)

TFT is a powerful multi-horizon time series forecasting DL model. It is a novel attention-based architecture with specific input processing to capture complex temporal patterns, while its interpretability features allow users to understand which factors drive the predictions. TFT excels in processing diverse inputs and making accurate multi-step forecasts.

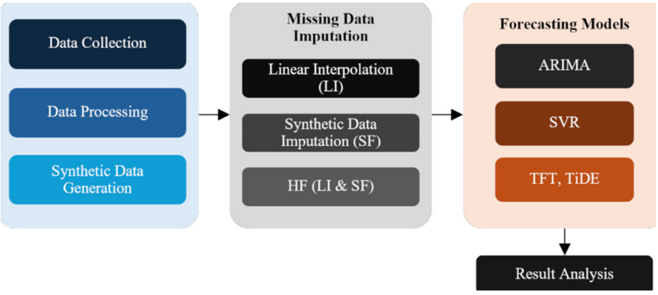

Fig. 3. Workflow of the study.

F. Performance Evaluation

MAE was used to evaluate the performance of the imputation methods and to measure the forecasting accuracy of the models implemented on the datasets. MAE is a commonly used metric to evaluate the accuracy of a model. It calculates the average absolute difference between the actual and forecast values without considering their directions. It is nonsymmetric and does not penalize large errors. All errors are equally weighted, as shown in (3).

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y_i}| \qquad (3)$$

Figure 3 illustrates the workflow, which consists of three sections, the data collection and processing phase, various missing data imputation processes, and the implementation of forecasting models on the datasets.

IV. EXPERIMENTAL RESULTS

This section evaluates the performance of the datasets using the various models over multiple forecasting horizons. Tables I and II show the overall performance of the models implemented to compare the forecasting accuracy over the three datasets.

A. Performance Analysis of the Imputation Methods

Table I shows the forecasting results obtained on LI, SF, and the proposed HF method using SVR and ARIMA, and Table II represents the MAE scores of the TFT and TiDE models. For ARIMA, the best MAE accuracy was obtained using the HF method, followed by SF. Table I shows that for all forecasting horizons, ARIMA on the proposed HF imputed dataset achieved the best MAE. MV imputation using the SF method provided better accuracy than the LI imputation method, except for the 24-hour forecasting horizon (MAE of 5.08 compared to 5.07 for LI). For the SVR model, the best MAE was obtained using SF, which outperformed HF and LI in all forecasting horizons. The LI method achieved better forecasting than the HF method in forecasting horizons of 24, 48, and 160 hours. The HF method outperformed the LI and SF methods in all models except SVR.

TABLE I. MAE SCORES OF ARIMA AND SVR ON LI, SF, AND HF DATASETS

Dataset	LI	SF	HF	AF	SF	HF
Model	ARIMA			SVR		
Forecast length (hrs)	MAE (kWh)					
24	5.07	5.08	4.93	1.00	0.97	1.04
48	4.85	4.76	4.69	2.42	2.17	2.47
72	5.25	5.24	5.06	3.01	2.64	2.99
120	4.99	4.89	4.81	3.58	3.20	3.55
168	4.58	4.53	4.49	3.38	3.13	3.43

TABLE II. MAE SCORE OF TFT AND TIDE ON LI, SF, AND HF DATASETS

Dataset	LI	SF	HF	AF	SF	HF
Model	TFT			TiDE		
Forecast length (hrs)	MAE (kWh)					
24	0.89	1.33	0.79	1.25	1.54	0.92
48	1.77	1.68	1.73	1.98	2.84	1.47
72	1.99	2.00	1.78	2.21	2.64	1.69
120	2.92	2.62	2.44	2.28	2.57	1.92
168	2.25	2.14	1.85	1.88	2.14	1.80

In the case of the TFT model, HF outperformed both the LI and SF methods in all forecasting horizons, except for the 48-hour horizon where the SF dataset beat HF (MAE score of 1.68, compared to 1.73). For the TFT model, the second-best performance score was achieved using SF, where it beat the LI method in three forecasting horizons. Finally, in the case of the TiDE model, the proposed HF method achieved better results than the LI and SF imputation methods in all forecasting horizons. The next-best results were achieved by the LI method, which outperformed SF in all forecasting horizons as well.

B. Performance Analysis Based on the Models Implemented

This section discusses the forecasting accuracies obtained by the models for the various forecasting horizons on the three datasets. For the 24-hour forecast horizon, TFT outperformed the other models with an MAE of 0.79 on the HF dataset. Figure 4 shows the 24-hour load forecast for all models on the HF dataset. The TFT and TiDE models were closest to the actual data for the 48-hour prediction, while the farthest was ARIMA. TiDE produced the best MAE of 1.47, again on the proposed HF dataset, as shown in Figure 5. For the remaining forecast horizons, TiDE outperformed the other models in terms of forecast accuracy, as shown in Figures 6, 7, and 8. TiDE is specifically developed for long-term forecasting and outperforms the other models as the forecast length increases. However, the best forecast accuracies were all achieved on the dataset with the HF imputation method. The forecast results of all models strongly indicate that the proposed hybrid method is viable to impute MVs.

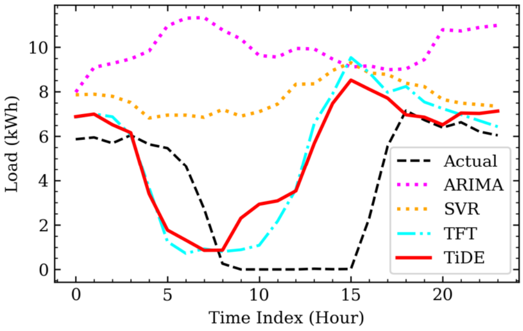

Fig. 4. MAE score of the forecasting models on 24-hour prediction length on the HF dataset.

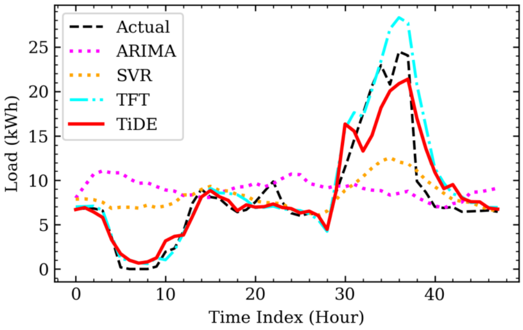

Fig. 5. MAE score of the forecasting models on 48-hours prediction length on the HF dataset.

When comparing the statistical, ML and DL models for all forecasting horizons, the DL models outperformed the others. However, for all models, the performance improved when the missing data imputation was performed using the proposed HF method, which outperformed both the SF and LI methods, except for the SVR model, which achieved its best performance on the SF dataset. Another observation is that as the length of the forecast horizon increases, the forecasting accuracy of the models also decreases for all the models implemented. The improvement in forecasting accuracy when imputing MVs using the hybrid approach over LI and SF strongly indicates that the proposed method is effective for imputing MVs in electricity load data.

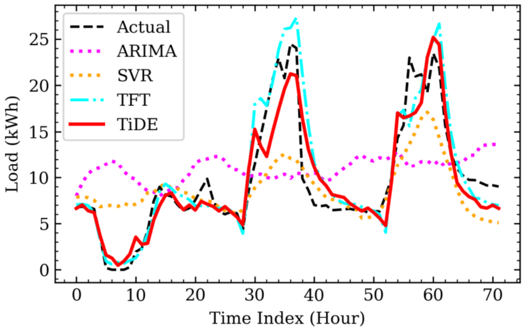

Fig. 6. MAE score of the forecasting models on 72-hours prediction length on the HF dataset.

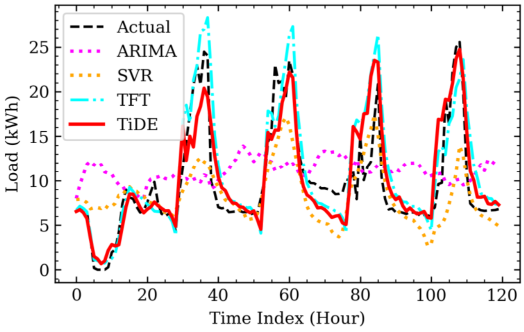

Fig. 7. MAE score of the forecasting models on 120-hours prediction length on the HF dataset.

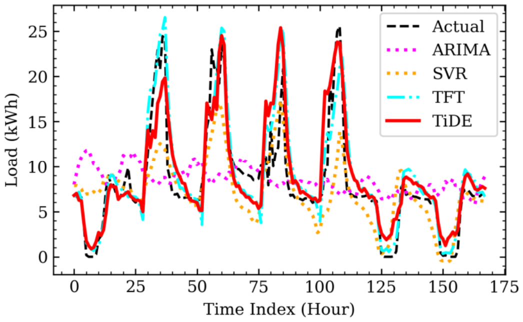

Fig. 8. MAE score of the forecasting models on 168-hours prediction length on the HF dataset.

V. CONCLUSIONS

This study focused on implementing synthetically generated data to fill in missing values in an electricity load dataset. LI, SF, and the proposed HF methods were employed to impute MVs. The experiments carried out with various models show that the performance of the statistical and DL models employed was improved when using the proposed hybrid imputation method. From the three datasets, HF improved the forecasting accuracy of ARIMA, TFT, and TiDE, whereas the SF dataset improved the performance of the SVR model. TiDE was the best-performing model for the overall forecasting horizons, outperforming the other models. Therefore, it can be concluded that, for time series datasets with many MVs with various missing block types, hybrid-filled data can be used to impute MVs and significantly increase the accuracy of forecasting models. This study suggests using the LI method for short MVs (2-3 hours) and synthetic data for larger continuous missing blocks.

ACKNOWLEDGMENTS

This research is part of a Ph.D. dissertation conducted at Naresuan University, Thailand, under a scholarship provided by His Majesty the King of Bhutan in collaboration with Naresuan University. This work was partially supported by the Global and Frontier Research University Fund, Naresuan University, under Grant R2567C002. The authors wish to thank Roy I. Morien and Reggie Dalman Hinoguin of the Naresuan University Graduate School for editing this manuscript's grammar, syntax, and general English expression.

REFERENCES

[1] S. K. Filipova-Petrakieva and V. Dochev, "Short-Term Forecasting of Hourly Electricity Power Demand: Reggresion and Cluster Methods for Short-Term Prognosis," *Engineering, Technology & Applied Science Research*, vol. 12, no. 2, pp. 8374–8381, Apr. 2022, https://doi.org/10.48084/etasr.4787.

[2] H. Nguyen and C. K. Hansen, "Short-term electricity load forecasting with Time Series Analysis," in *2017 IEEE International Conference on Prognostics and Health Management (ICPHM)*, Dallas, TX, USA, Jun. 2017, pp. 214–221, https://doi.org/10.1109/ICPHM.2017.7998331.

[3] L. Baur, K. Ditschuneit, M. Schambach, C. Kaymakci, T. Wollmann, and A. Sauer, "Explainability and Interpretability in Electric Load Forecasting Using Machine Learning Techniques – A Review," *Energy and AI*, vol. 16, May 2024, Art. no. 100358, https://doi.org/10.1016/j.egyai.2024.100358.

[4] S. Jung, J. Moon, S. Park, S. Rho, S. W. Baik, and E. Hwang, "Bagging Ensemble of Multilayer Perceptrons for Missing Electricity Consumption Data Imputation," *Sensors*, vol. 20, no. 6, Jan. 2020, Art. no. 1772, https://doi.org/10.3390/s20061772.

[5] N. Ahmad, Y. Ghadi, M. Adnan, and M. Ali, "Load Forecasting Techniques for Power System: Research Challenges and Survey," *IEEE Access*, vol. 10, pp. 71054–71090, 2022, https://doi.org/10.1109/ACCESS.2022.3187839.

[6] M. H. Bin Kamilin and S. Yamaguchi, "Resilient Electricity Load Forecasting Network with Collective Intelligence Predictor for Smart Cities," *Electronics*, vol. 13, no. 4, Jan. 2024, Art. no. 718, https://doi.org/10.3390/electronics13040718.

[7] A. R. Munappy, J. Bosch, H. H. Olsson, A. Arpteg, and B. Brinne, "Data management for production quality deep learning models: Challenges and solutions," *Journal of Systems and Software*, vol. 191, Sep. 2022, Art. no. 111359, https://doi.org/10.1016/j.jss.2022.111359.

[8] J. Jeong, T. Y. Ku, and W. K. Park, "Denoising Masked Autoencoder-Based Missing Imputation within Constrained Environments for Electric Load Data," *Energies*, vol. 16, no. 24, Jan. 2023, Art. no. 7933, https://doi.org/10.3390/en16247933.

[9] G. R. Hemanth and S. Charles Raja, "Proposing suitable data imputation methods by adopting a Stage wise approach for various classes of smart meters missing data – Practical approach," *Expert Systems with Applications*, vol. 187, p. 115911, Jan. 2022, https://doi.org/10.1016/j.eswa.2021.115911.

[10] J. Zhang and P. Yin, "Multivariate Time Series Missing Data Imputation Using Recurrent Denoising Autoencoder," in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, San Diego, CA, USA, Nov. 2019, pp. 760–764, https://doi.org/10.1109/BIBM47256.2019.8982996.

[11] S. N. Hussain, A. Abd Aziz, M. J. Hossen, N. A. Ab Aziz, G. R. Murthy, and F. Bin Mustakim, "A Novel Framework Based on Cnn-Lstm Neural Network for Prediction of Missing Values in Electricity Consumption Time-Series Datasets," *Journal of Information Processing Systems*, vol. 18, no. 1, pp. 115–129, 2022, https://doi.org/10.3745/JIPS.04.0235.

[12] J. Hwang and D. Suh, "CC-GAIN: Clustering and classification-based generative adversarial imputation network for missing electricity consumption data imputation," *Expert Systems with Applications*, vol. 255, Dec. 2024, Art. no. 124507, https://doi.org/10.1016/j.eswa.2024.124507.

[13] X. Shen, H. Zhao, Y. Xiang, P. Lan, and J. Liu, "Short-term electric vehicles charging load forecasting based on deep learning in low-quality data environments," *Electric Power Systems Research*, vol. 212, Nov. 2022, Art. no. 108247, https://doi.org/10.1016/j.epsr.2022.108247.

[14] H. Demirhan and Z. Renwick, "Missing value imputation for short to mid-term horizontal solar irradiance data," *Applied Energy*, vol. 225, pp. 998–1012, Sep. 2018, https://doi.org/10.1016/j.apenergy.2018.05.054.

[15] Y. Li *et al.*, "Load Profile Inpainting for Missing Load Data Restoration and Baseline Estimation," *IEEE Transactions on Smart Grid*, vol. 15, no. 2, pp. 2251–2260, Mar. 2024, https://doi.org/10.1109/TSG.2023.3293188.

[16] B. Cho *et al.*, "Effective Missing Value Imputation Methods for Building Monitoring Data," in *2020 IEEE International Conference on Big Data (Big Data)*, Atlanta, GA, USA, Dec. 2020, pp. 2866–2875, https://doi.org/10.1109/BigData50022.2020.9378230.

[17] J. Yoon, D. Jarrett, and M. van der Schaar, "Time-series generative adversarial networks," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Vancouver, Canada, Sep. 2019, pp. 5508–5518.

[18] J. Peppanen, Xiaochen Zhang, S. Grijalva, and M. J. Reno, "Handling bad or missing smart meter data through advanced data imputation," in *2016 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT)*, Minneapolis, MN, USA, Sep. 2016, pp. 1–5, https://doi.org/10.1109/ISGT.2016.7781213.

[19] Y. Mao, M. Yang, P. Li, and Z. Ou, "A Missing Data Imputation Method for Electricity Consumption Data Based on TCN-Attention with Mask Tokens," in *2024 4th International Conference on Consumer Electronics and Computer Engineering (ICCECE)*, Guangzhou, China, Jan. 2024, pp. 513–517, https://doi.org/10.1109/ICCECE61317.2024.10504227.

[20] M. N. Noor, A. S. Yahaya, N. A. Ramli, and A. M. M. A. Bakri, "Filling Missing Data Using Interpolation Methods: Study on the Effect of Fitting Distribution," *Key Engineering Materials*, vol. 594–595, pp. 889–895, 2014, https://doi.org/10.4028/www.scientific.net/KEM.594-595.889.

[21] I. Goodfellow *et al.*, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, Jul. 2020, https://doi.org/10.1145/3422622.

[22] C. Wang, Y. Cao, S. Zhang, and T. Ling, "A Reconstruction Method for Missing Data in Power System Measurement Based on LSGAN," *Frontiers in Energy Research*, vol. 9, Mar. 2021, https://doi.org/10.3389/fenrg.2021.651807.

[23] Z. Chang, S. Liu, Z. Cai, and G. Tu, "ANODE-GAN: Incomplete Time Series Imputation by Augmented Neural ODE-Based Generative Adversarial Networks," in *Artificial Neural Networks and Machine Learning – ICANN 2023*, Heraklion, Greece, 2023, pp. 16–27, https://doi.org/10.1007/978-3-031-44192-9_2.

[24] S. Aissa and K. M. Tarek, "Time Generative adversarial network for the generation of electricity load data," in *2023 International Conference on Control, Automation and Diagnosis (ICCAD)*, Rome, Italy, May 2023, pp. 1–5, https://doi.org/10.1109/ICCAD57653.2023.10152457.

[25] A. Kammoun, R. Slama, H. Tabia, T. Ouni, and M. Abid, "Generative Adversarial Networks for face generation: A survey," *ACM Computing Surveys*, Mar. 2022, https://doi.org/10.1145/1122445.1122456.

[26] Q. Wen *et al.*, "Time Series Data Augmentation for Deep Learning: A Survey," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, Montreal, Canada, Aug. 2021, pp. 4653–4660, https://doi.org/10.24963/ijcai.2021/631.

[27] Z. Yang, Y. Li, and G. Zhou, "TS-GAN: Time-series GAN for Sensor-based Health Data Augmentation," *ACM Transactions on Computing for Healthcare*, vol. 4, no. 2, pp. 12:1-12:21, Dec. 2023, https://doi.org/10.1145/3583593.

[28] D. S. Lee and S. Y. Son, "PV Forecasting Model Development and Impact Assessment via Imputation of Missing PV Power Data," *IEEE Access*, vol. 12, pp. 12843–12852, 2024, https://doi.org/10.1109/ACCESS.2024.3352038.

[29] E. Afrifa-Yamoah, U. A. Mueller, S. M. Taylor, and A. J. Fisher, "Missing data imputation of high-resolution temporal climate time series data," *Meteorological Applications*, vol. 27, no. 1, 2020, Art. no. e1873, https://doi.org/10.1002/met.1873.

[30] Z. Zhang, "Missing data imputation: focusing on single imputation," *Annals of translational medicine*, vol. 4, no. 1, Jan. 2016, https://doi.org/10.3978/j.issn.2305-5839.2015.12.38.