JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 28 No. 11, pp. 29-42, November 2023
https://doi.org/10.9708/jksci.2023.28.11.029

A Comparative Study on Data Augmentation Using Generative Models for Robust Solar Irradiance Prediction

Jinyeong Oh*, Jimin Lee*, Daesungjin Kim**, Bo-Young Kim**, Jihoon Moon*

*Student, Dept. of AI and Big Data, Soonchunhyang University, Asan, Korea
*Student, Dept. of AI and Big Data, Soonchunhyang University, Asan, Korea
**Student, Asan Middle School, Asan, Korea
**Teacher, Asan Middle School, Asan, Korea
*Professor, Dept. of AI and Big Data, Soonchunhyang University, Asan, Korea

[Abstract]

In this paper, we propose a method to enhance the prediction accuracy of solar irradiance for three major South Korean cities: Seoul, Busan, and Incheon. Our method entails the development of five generative models—vanilla GAN, CTGAN, Copula GAN, WGANGP, and TVAE—to generate independent variables that mimic the patterns of existing training data. To mitigate the bias in model training, we derive values for the dependent variables using random forests and deep neural networks, enriching the training datasets. These datasets are integrated with existing data to form comprehensive solar irradiance prediction models. The experimentation revealed that the augmented datasets led to significantly improved model performance compared to those trained solely on the original data. Specifically, CTGAN showed outstanding results due to its sophisticated mechanism for handling the intricacies of multivariate data relationships, ensuring that the generated data are diverse and closely aligned with the real-world variability of solar irradiance. The proposed method is expected to address the issue of data scarcity by augmenting the training data with high-quality synthetic data, thereby contributing to the operation of solar power systems for sustainable development.

▸ **Key words**: Solar Energy Forecasting, Generative Data Augmentation, Deep Learning Models, Environmental Sustainability, Data Insufficiency Solutions

• First Author: Jinyeong Oh, Corresponding Author: Jihoon Moon
*Jinyeong Oh (wlsdud3523@sch.ac.kr), Dept. of AI and Big Data, Soonchunhyang University
*Jimin Lee (dlwlals7359@sch.ac.kr), Dept. of AI and Big Data, Soonchunhyang University
**Daesungjin Kim (kimjc651207@naver.com), Asan Middle School
**Bo-Young Kim (bboya414@naver.com), Asan Middle School
*Jihoon Moon (jmoon22@sch.ac.kr), Dept. of AI and Big Data, Soonchunhyang University
• Received: 2023. 10. 18, Revised: 2023. 11. 13, Accepted: 2023. 11. 13.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

[요 약]

본 논문은 서울, 부산, 인천과 같은 대한민국의 주요 도시들을 대상으로 일사량 예측 정확도를 향상하기 위한 방법론을 제안한다. 제안한 방법론은 먼저 GAN, CTGAN, Copula GAN, WGANGP, TVAE 등 다섯 가지 생성 모델을 이용하여 기존 학습 데이터와 유사한 독립 변수들을 생성한다. 다음으로 모델 학습에서의 데이터 편향성을 개선하고자, 생성한 독립 변수들에서 각각 랜덤 포레스트와 심층 신경망을 통해 종속 변숫값을 도출하여 학습 데이터 셋을 구축하고, 이를 기존 학습 데이터 셋과 결합하여 예측 모델을 구성한다. 실험 결과, 증강된 데이터 셋으로 학습한 모델들은 기존 데이터 셋으로 학습한 모델들보다 향상된 성능을 나타내었다. 특히 CTGAN은 복잡한 다변량 데이터 관계를 효과적으로 다루는 메커니즘으로 인해 우수한 결과를 도출하였으며, 생성된 데이터는 일사량의 다양한 변화와 실제 변동성과 효과적으로 반영하였다. 제안한 방법론은 고품질의 생성 데이터로 학습 데이터를 증강함으로써, 데이터 부족 현상 문제를 다룰 수 있을 뿐만 아니라 지속 가능한 발전을 위한 태양광 발전 시스템 운영에도 이바지할 수 있을 것으로 기대한다.

▸ **주제어**: 태양 에너지 예측, 생성 데이터 증강, 딥러닝 모델, 환경 지속 가능성, 데이터 부족 솔루션

I. Introduction

In our globally interconnected society, a reliable energy and power supply has become an essential cornerstone for maintaining consistent economic growth. The rapid acceleration of urbanization and industrialization across the globe is driving an unprecedented surge in energy demand. Historically, fossil fuels have been the primary source to meet this demand, but the global energy market is now undergoing a significant transformation. Several factors have prompted this shift. Firstly, reserves of fossil fuels are finite and dwindling. Secondly, considerable environmental implications are linked to their use, such as greenhouse gas emissions leading to climate change. Lastly, substantial economic challenges are associated with the extraction, transportation, and usage of these fuels. These constraints have catalyzed a transition towards renewable energy sources that are more sustainable and economically viable in the long run [1]. Solar energy is emerging as a leading player in this renewable revolution due to its eco-friendly characteristics and cost-effectiveness, which have garnered considerable global attention [2]. South Korea is one nation that has been particularly proactive in adopting solar power within its energy portfolio. By 2020, South Korea had successfully established solar power facilities with an impressive capacity of 5.5 GW. This achievement underscores South Korea's commitment to renewable energy adoption and signals its recognition of solar power's crucial role in future energy security [3]. This continued investment in solar infrastructure suggests a promising trajectory for growth within South Korea's solar sector.

The essence of solar power systems lies in their ability to transform sunlight into electricity using solar panels. However, the generation of solar energy has its challenges. Predominantly, the output is susceptible to fluctuations driven by various environmental factors, rendering its prediction complex and demanding cutting-edge technical approaches [4]. One primary determinant of solar energy production is solar irradiance [5]. Thus, ensuring precise irradiance forecasting is not merely a technical challenge but a necessity. Such accuracy bolsters the efficiency of solar energy systems and fortifies their stability, emphasizing the crucial role of irradiance prediction in the overarching framework of renewable energy

research [6]. Given its critical implications for energy production, solar irradiance prediction has been a focal point of numerous research endeavors. Traditionally, these predictions have been anchored in time series analysis, leveraging historical data on temperature, humidity, and wind speed as proxies for environmental variables [7]. Conventional methods often use statistical approaches or models such as the autoregressive integrated moving average (ARIMA) to analyze these time-series data. However, these traditional techniques must be revised when grappling with environmental variables' non-linear and intricate correlations [8].

This realization has ushered in a paradigm shift towards artificial intelligence-based methodologies. Machine learning (ML) and, more prominently, deep learning (DL) technologies have taken center stage in recent irradiance prediction efforts, demonstrating their prowess in decoding complex time-series data patterns [9, 10]. Deep learning techniques consistently showcase superior prediction performance, especially when leveraging architectures such as ensemble learning and hybrid models [10]. Voyant et al. [9] undertook a comparative study on machine learning-based approaches for solar radiation predictions, pitting algorithms such as artificial neural networks (ANNs) and ARIMA against each other. Their findings underscored the exemplary prediction performance of support vector machines (SVM) and random forest (RF) algorithms. Similarly, Lee et al. [11] juxtaposed the performance of four ensemble learning models for irradiance prediction, noting that ensemble methods outperformed standalone models, reducing prediction variance. A study by Kumari et al. [12] weighed the prediction prowess of various deep learning architectures, concluding that hybrid deep learning models offer optimal performance. Yan et al. [13] innovatively combined the gated recurrent unit (GRU) with attention mechanisms, crafting a hybrid deep learning model adept at seasonally adjusted short-term solar irradiance prediction, surpassing conventional single-model predictions in accuracy.

After exploring various models and techniques, it becomes evident that the core challenge to achieving superior solar irradiance predictions lies in the available data's abundance and quality. Traditional approaches in the field have predominantly hinged on leveraging extensive data repositories, often spanning extensive durations, to ensure reliable short-term and long-term irradiance forecasting [5, 7, 8]. Despite these efforts, a significant research gap exists in ensuring optimal prediction accuracy when the available training data is scanty or insufficient [14]. In response to this challenge, the rise of generative models, especially generative adversarial networks (GANs) and their variants, offers a beacon of hope. These models have been developed to generate data that closely resembles genuine datasets, thus compensating for limited training data and paving the way for more robust prediction algorithms [15]. Figure 1 elucidates this concept further by presenting a schematic overview of a foundational generative model's architecture. Their unparalleled ability to produce realistic data has led generative models to find applications in diverse domains, from image processing and computer vision to natural language processing. Moreover, our current investigation underscores their immense promise in time-series analysis, particularly for refining solar irradiance forecasting in scenarios with limited data.

The deployment of GAN models for time-series predictions has become increasingly prominent. Improved forecasting performances have been demonstrated by generating supplementary data through these generative models or by integrating GAN with other deep learning architectures. Huang et al. [16] proposed a solar power generation prediction using conditional GANs (CGAN), continually refining prediction accuracy by juxtaposing real and predicted values through bidirectional long short-term memory (Bi-LSTM). Li

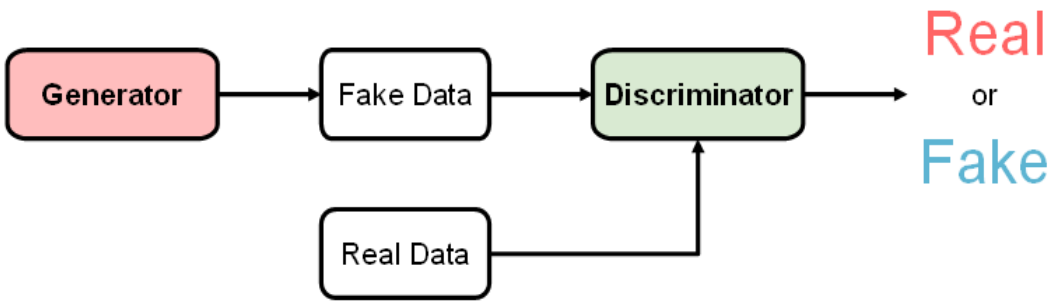

Fig. 1. Schematic Diagram of Vanilla GAN's Structure

et al. [17] combined the strengths of Wasserstein GAN (WGAN) and long short-term memory (LSTM) after segmenting solar irradiance output signals into multiple sub-sequences, integrating individual component predictions for the final output. Moon et al. [18] proffered a two-tiered data generation approach for enhancing electric load forecasting and amalgamating generated and authentic datasets for model perfection. Wang et al. [19] used GANs to bolster weather data for solar power generation predictions, training a convolutional neural network (CNN)-based classifier on original and generated datasets. Fekri et al. [20] adopted the recurrent GAN (R-GAN) for generating energy consumption data for smart grid operations, validating that GAN-based data can train energy prediction models. Tian et al. [21] anchored their building energy consumption predictions on parallel learning theories using GANs, showing that hybrid datasets lead to superior forecasting outcomes.

Drawing inspiration from these studies, we undertake a comparative exploration of five unique generative models to bolster limited training datasets. A hallmark of our research is using genuine solar irradiance data from three major South Korean metropolitan regions: Seoul, Busan, and Incheon. Our proposed methodology intertwines generative models to refine training datasets, subsequently deploying RFs and deep neural networks (DNNs) for rigorous solar irradiance prediction. This method's efficacy is corroborated through stringent evaluations using root mean square error (RMSE) and mean absolute error (MAE) metrics. Ultimately, our findings promise to be a beacon for the energy and data science sectors, emphasizing the quintessential role of generative models in tackling data inadequacies and fortifying prediction accuracy.

The structure of this paper is meticulously organized as follows: Section 2 elucidates the methodologies proposed in this paper. Section 3 offers a rigorous analysis of the predictive performance achieved in our research. Section 4 draws the study to a close by presenting the conclusions, highlighting limitations, and suggesting future research trajectories.

II. Methodology

2.1 Data Collection and Preparation

To emulate real-world scenarios of data scarcity, our study utilized solar irradiance data from three major metropolitan cities in South Korea: Seoul, Busan, and Incheon [5]. The datasets encompass the period from January 1, 2019, to December 31, 2020. For training and validation, data from 2019 served as the training set, while the entirety of the 2020 data was earmarked for testing. Table 1 provides a detailed breakdown of the features encapsulated in our datasets. These variables have been carefully chosen to encapsulate temporal, climatic, and solar attributes influencing irradiance levels.

Table 1. Information on Dataset Features

Variables	Description	Variable Type
$Month_{sin}$ (Input)	Sine Value of Month	Continuous on [-1, 1]
$Month_{cos}$ (Input)	Cosine Value of Month	Continuous on [-1, 1]
Day_{sin} (Input)	Sine Value of Day	Continuous on [-1, 1]
Day_{cos} (Input)	Cosine Value of Day	Continuous on [-1, 1]
$Hour_{sin}$ (Input)	Sine Value of Hour	Continuous on [-1, 1]
$Hour_{cos}$ (Input)	Cosine Value of Hour	Continuous on [-1, 1]
Temp (Input)	Temperature (in Degrees Celsius)	Continuous
Humi (Input)	Relative Humidity (in Percentage)	Continuous
WS (Input)	Wind Speed (in m/s)	Continuous
Solar (Output)	Solar Irradiance (in MJ/m^2)	Continuous

2.2 Model Proposition for Data Augmentation

In this research, we address the challenge of solar irradiance prediction under conditions of data insufficiency. We propose a novel approach that leverages the capabilities of generative models to supplement the lack of training data and enhance predictive efficiency. Our proposed model involves a two-step process, which is visually represented in Figure 2, illustrating these two integral steps of our model proposition:

- **Input Data Generation:** This step synthesizes new input data that replicates the characteristics and length of the original dataset. It is not a mere duplication; the generative models are trained to understand the intrinsic patterns and variations in the original data and then reproduce those nuances in the generated data. This process ensures that our augmented data maintains the richness of information while providing additional diversity for training.
- **Output Data Construction:** After synthesizing the input data, we employ a pre-trained regression model, which has been initially trained on the original dataset. We obtain the corresponding output data by feeding the generated input data into this model. This step ensures that our synthetic data pairs (both input and output) maintain a semblance to real-world data scenarios.

The efficacy of our data augmentation relies significantly on the capabilities of the chosen generative models. Here is a comprehensive understanding:

- **Vanilla GAN:** A foundational model in the GAN family, Vanilla GAN is a trailblazer in generative modeling. Recognized for consistently generating high-quality data, its primary strength lies in establishing a benchmark for other advanced models. Given its simplicity yet robust performance, it is the litmus test against which other GANs are often evaluated.
- **Conditional tabular GAN (CTGAN):** Building on the core GAN architecture, CTGAN [22] is tailored for tabular datasets. It employs

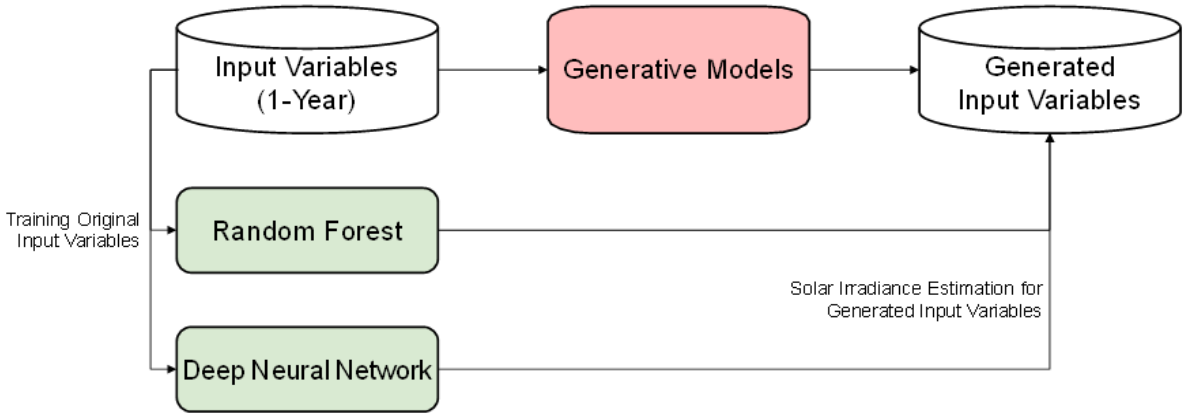

Fig. 2. Process Flow of Data Augmentation Using Generative Models

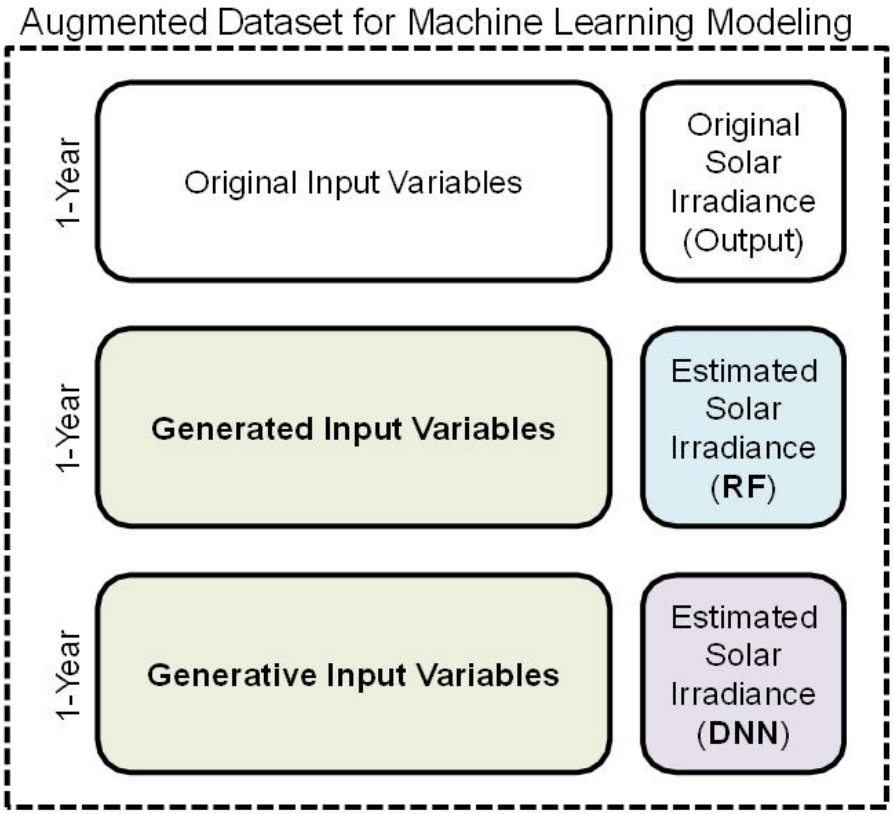

Fig. 3. Structure of the Proposed Enhanced Data Format

quantification mechanisms to interpret and generate structured data, addressing the challenges of non-sequential patterns making it ideal for datasets needing detailed pattern recognition.

- **Copula GAN:** Expanding on standard data generation techniques, Copula GAN [23] incorporates the Copula function, which excels at identifying relationships between multivariate distributions. Mimicking data distribution creates a genuine-seeming correlation among generated data points, ensuring the output appears authentic and preserves realistic interactions between variables.
- **Wasserstein GAN with Gradient Penalty (WGANGP):** Evolving from WGAN, the WGANGP [24] incorporates a gradient penalty. This addition aids the training by adjusting the model's parameters whenever the discriminator finds distinguishing real from generated data challenging, ensuring sustained data quality throughout iterations.
- **Temporal Variational Autoencoder (TVAE):** Specifically designed for time-series data, TVAE [25] encapsulates the nuances of temporal patterns. Unlike standard autoencoders, it considers the sequence, chronology, and time-bound correlations, making it a preferred choice when dealing with datasets where the sequence and time-based patterns play a pivotal role.

Effectively predicting time-series data often necessitates the availability of a comprehensive dataset. This research addresses this need by leveraging a generative model trained on one year's input data to reproduce a dataset of equivalent length. We employed two state-of-the-art regression techniques, RF and DNN, to deduce the solar irradiance corresponding to this synthetically generated input.

- **RF:** As an ensemble learning method, RF offers

a holistic approach that amalgamates the results of multiple decision trees to yield a more accurate and stable prediction [26]. It is particularly lauded for its ability to deliver reliable outcomes without requiring exhaustive hyperparameter tuning, distinguishing it as a versatile and effective algorithm for various prediction tasks [27].

- **DNN:** Contrasting traditional machine learning models, DNNs possess a deeper architecture, enabling them to model intricate relationships inherent in complex datasets [28]. The multi-layered nature of DNNs facilitates the capture of subtle patterns and dependencies, especially pivotal in time-series data where temporal correlations are prevalent [29].

Opting for a dual-regression strategy, which integrates both RF and DNN, was a deliberate and methodical decision underpinned by empirical evidence and the inherent strengths of each technique [30]. Combining predictions and features learned from RF and DNN ensures a well-rounded representation in our augmented dataset. Such a synthesis ensures that the strengths of one model can compensate for the limitations of the other, leading to a diverse and comprehensive dataset. This approach reinforces the reliability of our augmented dataset and makes it a more holistic training ground for subsequent predictive models. By synergizing the outputs of regression models trained on real-world data through this approach, our dataset burgeoned from an initial count of 4,015 data points to a formidable 12,045 for each region, as delineated in Figure 3. Such an expansive and enriched dataset lays solid groundwork, priming us for a more dependable and precise final prediction performance evaluation.

2.3 Solar Irradiance Prediction

To achieve an intricate and rigorous evaluation of predictive capabilities, we resorted to both RF and DNN models as embodied within the Orange3 framework. Following the generation of the augmented dataset—a fusion of real and synthetically generated data using RF and DNN—we trained both RF and DNN models again, exploiting the combined dataset's richness. The goal was to predict solar irradiance values for the remaining evaluation set.

III. Results

3.1 Implementation and Evaluation Framework

In pursuit of advanced solar irradiance forecasting, our experimental efforts unfolded within a Python 3.9.0 environment, utilizing the capabilities of TensorFlow 2.5.0 and PyTorch 1.10.2. These leading-edge deep learning frameworks were pivotal for the meticulous execution and optimization of our generative models, ensuring peak performance and scalability. Our methodology was expansive, integrating a suite of sophisticated generative architectures: Vanilla GAN, CTGAN, Copula GAN, WGANGP, and TVAE, each contributing to the goal of enhanced solar irradiance prediction.

Grasping the nuances of how generative models produce and refine synthetic data is critical. Such an understanding is pivotal, as it empowers professionals to employ synthetic data to reflect or supplement actual datasets accurately. The t-distributed stochastic neighbor embedding (T-SNE) visualizations serve as a vital interpretive mechanism, showcasing how each generative model captures the intricate patterns and variances intrinsic to solar irradiance data. By comparing the T-SNE plots of original and synthesized data, we can critically assess the models' performance in creating data that is not only statistically similar but also structurally representative of the complex dynamics observed in solar irradiance patterns. Such comparative analysis is indispensable for confirming that the augmented data maintains the multidimensional relationships in actual irradiance measurements. Our study highlights this process by

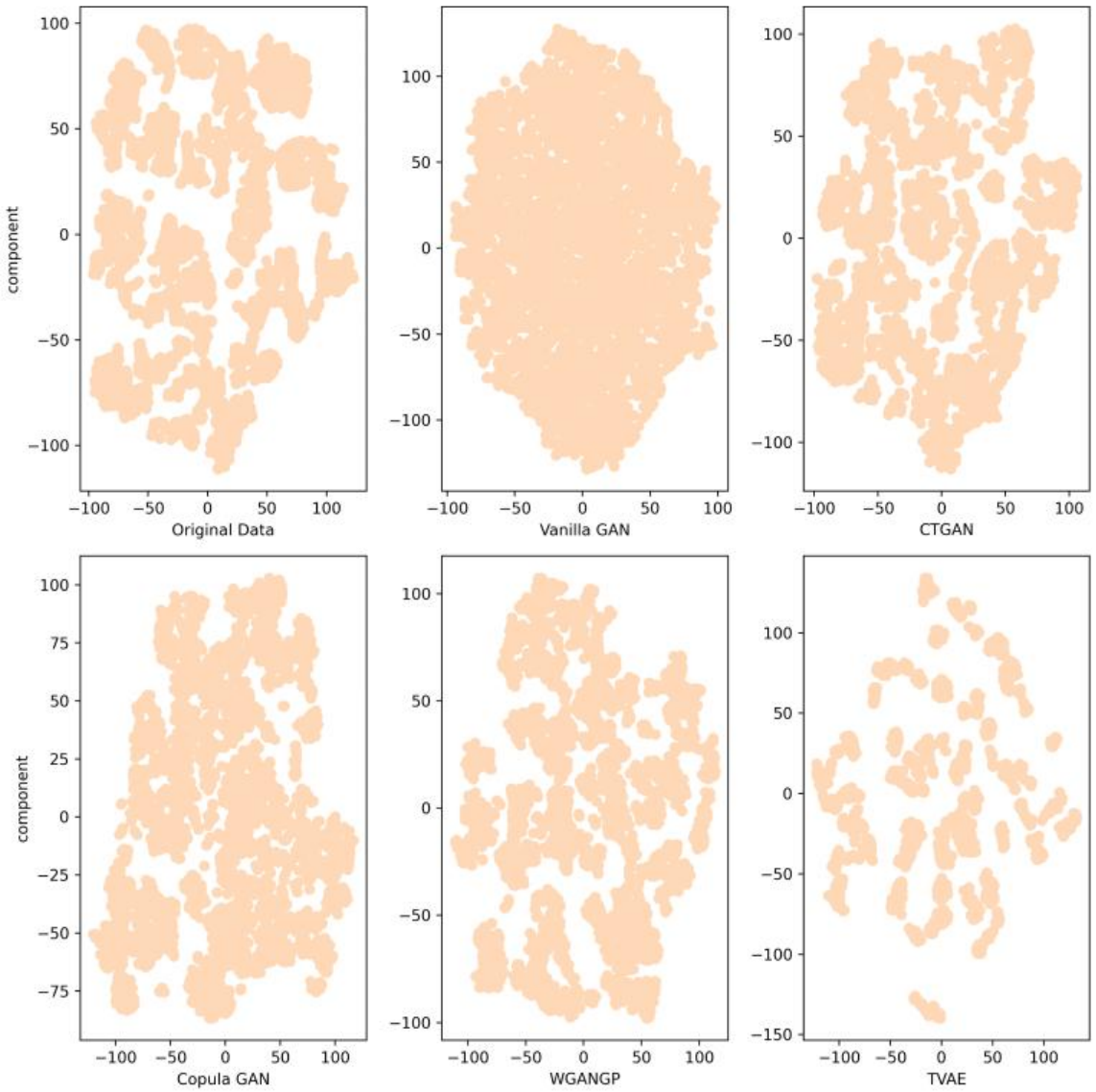

Fig. 4. T-SNE Visualization of Original and Augmented Data in Seoul

providing visualizations in Figures 4-6, utilizing T-SNE to distill complex data effectively.

The practical implications of generative models are underscored by interpreting these T-SNE visualizations. For instance, the synthetic data generated by the CTGAN model exhibits a high degree of similarity with the original dataset, indicating its proficiency in replicating intricate data patterns. This fidelity is crucial for applications where precise modeling of solar patterns is required. Conversely, the synthetic data from the GAN model shows a marked distinction from the original dataset in T-SNE visualizations, suggesting a potential shortfall in capturing complex data relationships. Such discrepancies could harm the model's effectiveness in accurately forecasting solar irradiance.

The Copula GAN and WGANGP models present synthetic data that, while not a perfect match, resembles the original data. This level of approximation could be beneficial in situations where an exact replication of the data is unnecessary. However, a general representation is still needed, such as in the early stages of model training. Meanwhile, the synthetic data from the TVAE model tends to cluster, implying a

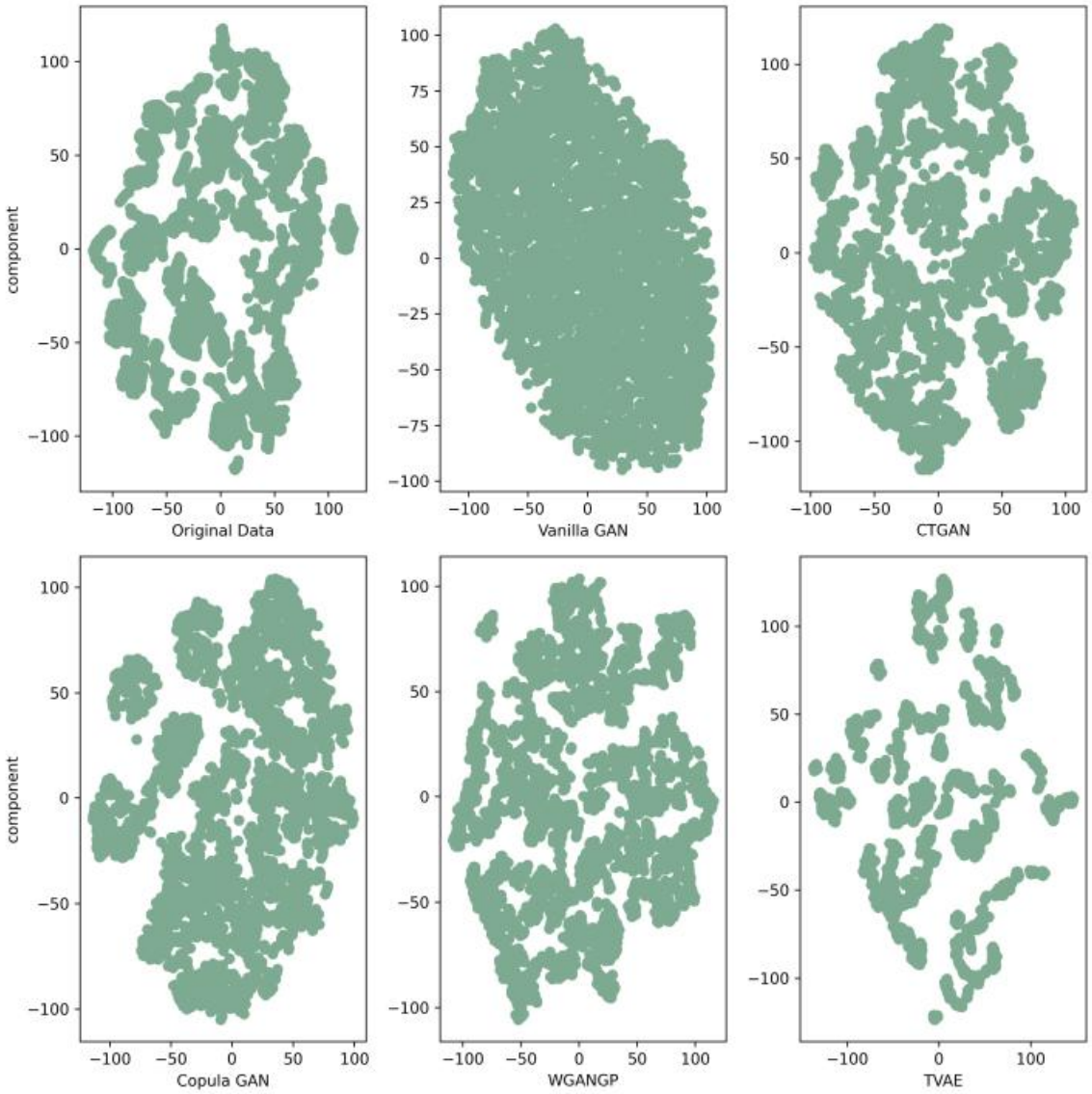

Fig. 5. T-SNE Visualization of Original and Augmented Data in Busan

concentration on particular aspects of the dataset. This characteristic could be advantageous when the objective is to model specific phenomena within the solar irradiance field.

The selection of an appropriate generative model hinges on the particular requirements of the task. The CTGAN model is suitable for cases where high-fidelity replication is paramount. The WGANGP or Copula GAN models are apt choices when the overarching data trend is more important than granular details. The TVAE model emerges as the preferred option for targeting specific features within a dataset. Ultimately, the insights gained from these visualizations inform the selection and application of generative models in solar irradiance forecasting. They ensure that the models employed are theoretically sound and capable of generating data that enhances the accuracy and reliability of solar irradiance predictions in real-world scenarios.

3.2 Augmented Data Prediction Performance

We harnessed the RF and DNN models for the subsequent predictive performance evaluation, both implemented within the Orange3 framework. Orange3, an open-source powerhouse, streamlines numerous tasks, from data mining to machine learning and even intricate tasks such as image

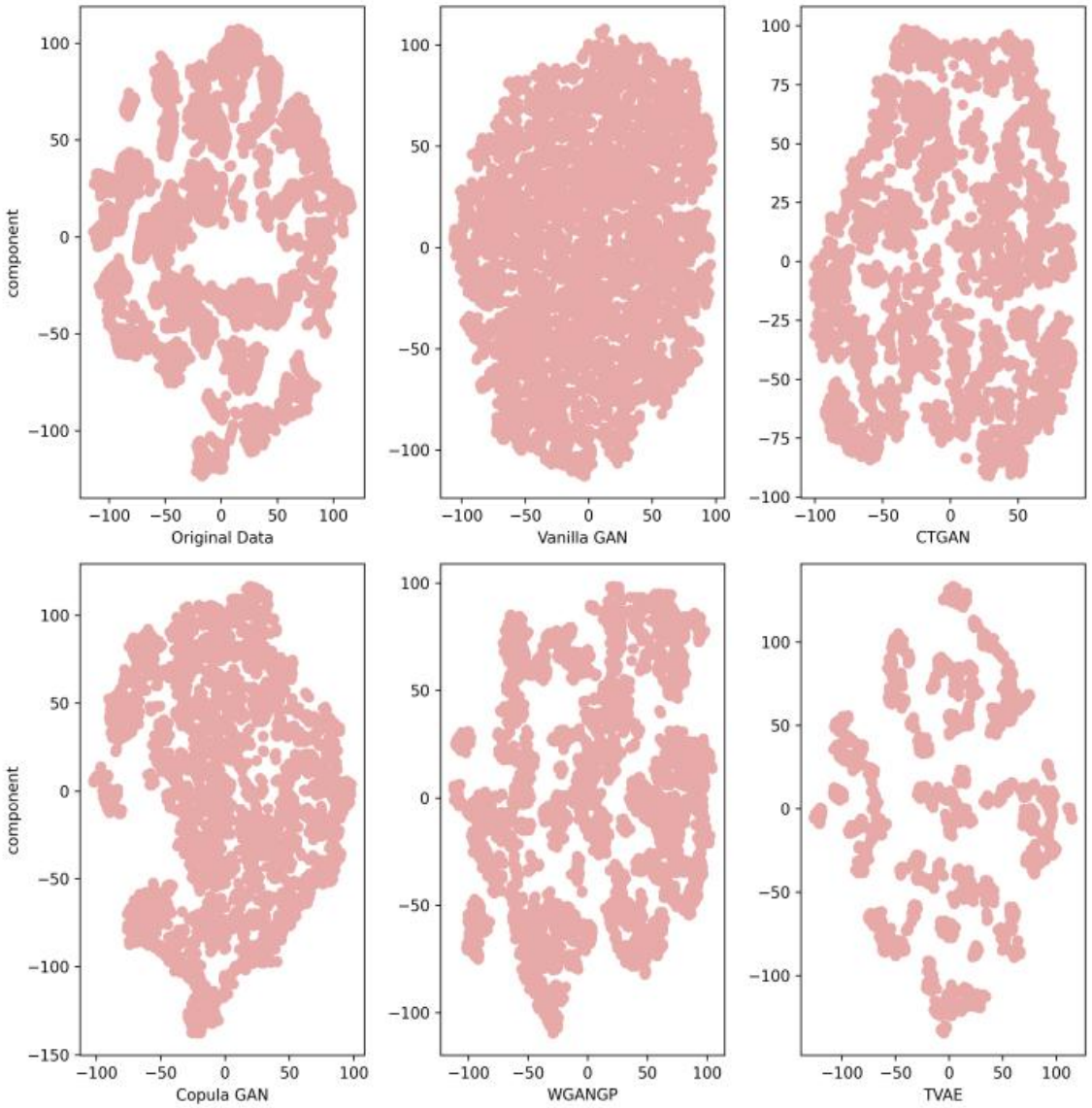

Fig. 6. T-SNE Visualization of Original and Augmented Data in Incheon

and time-series analysis. Its intuitive, no-coding interface benefits non-programmers and students eager for tangible data analytics experiences [31]. Among the algorithms at our disposal, the RF, an ensemble learning method, stands out, offering swift learning rates and impressive performance metrics across various applications. The evaluation of the predictive capabilities of our models relied heavily on two robust metrics: RMSE and MAE.

In predictive modeling, the quality and quantity of training data play a pivotal role in determining the accuracy and reliability of predictions. Our investigation aimed to refine the solar irradiance prediction performance across three critical regions: Seoul, Busan, and Incheon. Recognizing this, our study leveraged the capabilities of RF and DNN models to augment the original dataset. Specifically, after generating supplementary datasets using both RF and DNN, we integrated these enhanced datasets to form a comprehensive and enriched training set. This newly formed dataset, inherently more robust and encompassing than its predecessor, was subjected to further learning using both RF and DNN.

Table 2. MAE Comparison with Random Forest

Models	Seoul	Busan	Incheon
Original Data	0.772	0.807	0.692
Vanilla GAN	0.357	0.326	0.287
CTGAN	0.350	0.321	0.280
Copula GAN	0.355	0.329	0.282
WGANGP	0.352	0.333	0.289
TVAE	0.361	0.324	0.292

Table 3. RMSE Comparison with Random Forest

Models	Seoul	Busan	Incheon
Original Data	0.967	1.070	0.906
Vanilla GAN	0.494	0.535	0.403
CTGAN	0.491	0.535	0.398
Copula GAN	0.495	0.546	0.395
WGANGP	0.494	0.544	0.407
TVAE	0.505	0.538	0.416

The data in Tables 2 and 3 robustly quantify the RF model's improved predictive performance. Using MAE and RMSE as benchmarks, the model trained on data from generative models delivered more accurate predictions than the original data across Seoul, Busan, and Incheon. This finding is a testament to the enhanced capability of generative models in refining solar irradiance forecasting.

- **Table 2 (MAE):** For Seoul, data augmented using CTGAN registered the lowest MAE of 0.350, followed closely by WGANGP at 0.352, notably reducing the MAE from the original data's 0.772. Incheon and Busan followed a similar trend, with CTGAN and WGANGP often delivering the most refined predictions. This result underscores the transformative role of generative models in enhancing prediction accuracy.
- **Table 3 (RMSE):** When focusing on RMSE, a measure that gives a higher penalty to larger prediction errors, the CTGAN-augmented data for Seoul and Incheon yielded the lowest RMSE values of 0.491 and 0.395, respectively. Both Vanilla GAN and CTGAN achieved an RMSE of 0.535 for Busan, marking a significant reduction from the original data's RMSE of 1.070.

Table 4. MAE Comparison with DNN

Models	Seoul	Busan	Incheon
Original Data	0.799	0.874	0.780
Vanilla GAN	0.379	0.357	0.320
CTGAN	0.369	0.323	0.298
Copula GAN	0.354	0.327	0.297
WGANGP	0.379	0.343	0.320
TVAE	0.358	0.332	0.308

Table 5. RMSE Comparison with DNN

Models	Seoul	Busan	Incheon
Original Data	0.951	1.043	0.924
Vanilla GAN	0.545	0.623	0.478
CTGAN	0.527	0.545	0.426
Copula GAN	0.507	0.560	0.429
WGANGP	0.550	0.588	0.470
TVAE	0.515	0.560	0.453

In a deeper architecture, Tables 4 and 5 detail the performance metrics of the DNN. The DNN model, similar to the RF model, showcased marked improvements when trained on augmented data.

- **Table 4 (MAE):** The MAE values for the DNN trained on Copula GAN-augmented data stood out for Seoul and Incheon, registering values of 0.354 and 0.297, respectively, substantial improvements over the original data's values of 0.799 and 0.780. In contrast, for Busan, the CTGAN-augmented data led to the lowest MAE value of 0.323.
- **Table 5 (RMSE):** For RMSE, the Copula GAN model was particularly effective for Seoul, bringing down the RMSE to 0.507 from the original 0.951. Incheon, too, saw its lowest RMSE, with CTGAN at 0.426. Busan showcased a more even distribution with minor variations in RMSE across the generative models, but CTGAN marginally led the pack with an RMSE of 0.545.

Upon thoroughly scrutinizing the presented tables and performance metrics, it becomes evident that models trained on meticulously augmented data using generative models consistently outperform those trained solely on the original dataset. CTGAN stands out, frequently delivering

the most accurate predictions. This detailed analysis underscores the importance and effectiveness of our proposed data augmentation and iterative training methodology. It highlights the transformative potential of combining generative models with robust predictive algorithms such as RF and DNN in solar irradiance prediction.

IV. Conclusions

In this study, we harnessed the power of generative models to address the challenge of limited training data, seeking to optimize the performance of predictive models. Deployed within a Python 3.9.0 environment and leveraging state-of-the-art deep learning frameworks such as TensorFlow 2.5.0 and PyTorch 1.10.2, our methodology capitalized on advanced generative architectures, including Vanilla GAN, CTGAN, Copula GAN, WGANGP, and TVAE. The utility of these models became evident in their capacity to produce high-quality and consistent data, reinforcing the transformative potential of our data augmentation strategy. A cornerstone of our methodology was the rigorous validation of the synthesized data. Employing T-SNE, renowned for its capability to retain high-dimensional data structures, we could ensure that our augmented data mirrored the inherent characteristics and distribution of the original datasets.

Our results, anchored on RF and DNN models, showcased a pronounced enhancement in predictive accuracy across pivotal regions: Seoul, Busan, and Incheon. The augmented datasets, integrated from both RF and DNN outputs, created a comprehensive training set, further empowering the predictive prowess of our models. A granular examination of our results, as depicted in Tables 2-5, underscores the significant advantages of leveraging augmented data. Notably, CTGAN frequently outperformed other generative models, particularly in its capacity to minimize MAE and RMSE values. Drawing from our findings, the marriage of generative models with predictive algorithms manifests as a game-changing strategy, especially in domains like solar irradiance prediction. This paradigm addresses the limitations of sparse training data and offers an innovative approach that can significantly influence business decision-making.

Our investigation into solar irradiance prediction using generative models within the Orange3 framework has yielded promising results. However, it is important to note that this study was constrained by computational resources, precluding advanced deep learning techniques and exploring transfer learning to enhance model generalizability. Consequently, we could not delve into the transferability of the models to diverse geographic locations, which is essential for establishing the models' versatility in different environmental contexts. Moreover, the analysis did not modify the internal mechanics of the generative models, which could have provided deeper insights into their predictive capabilities. Future studies should aim to apply these models across varied locales to confirm their applicability and to harness more powerful computing resources, allowing for the incorporation of sophisticated machine learning algorithms that could further refine the models' accuracy and reliability.

Nevertheless, the vast potential of our methodology extends beyond the realm of solar irradiance prediction. Its versatility can be tapped across diverse challenges in various sectors, fortifying its contribution to the evolving landscape of data science. Despite the significant strides made, our methodology has its limitations. Future studies might probe into potential edge cases where generative augmentation might falter. Exploring and refining other generative architectures tailored for specific prediction challenges can further broaden the horizons of this research domain. As we navigate this exciting juncture, a continued focus on refining and expanding our

methodological arsenal promises a future teeming with innovation and impactful solutions.

ACKNOWLEDGEMENT

This study was supported by the MSIT (Ministry of Science, ICT), Korea, under the National Program for Excellence in SW, supervised by the IITP (Institute of Information & communications Technology Planning & Evaluation) in 2021 (2021-0-01399).

REFERENCES

[1] Y. Han, J. Moon, S. Rho, and H. Chang, "A robust and explainable missing value imputation method for smart meter data," Journal of Society for e-Business Studies, Vol. 27, No. 3, pp. 21-43, August 2022. DOI: 10.7838/jsebs.2022.27.3.021

[2] L. Cozzi et al., "World energy outlook 2020," International Energy Agency: Paris, France, pp. 1-461, 2020.

[3] 2023 SEMI-ANNAL Report, https://www.ctis.re.kr/ko/downloadBbsFile.do?atchmnflNo=9673.

[4] A. M. Omer, "Energy, environment and sustainable development," Renewable and Sustainable Energy Reviews, Vol. 12, No. 9, pp. 2265-2300, December 2008. DOI: 10.1016/j.rser.2007.05.001

[5] J. Moon, Y. Han, H. Chang, and S. Rho, "Multistep-ahead solar irradiance forecasting for smart cities based on LSTM, Bi-LSTM, and GRU neural networks," Journal of Society for e-Business Studies, Vol. 27, No. 4, pp. 27-52, November 2022. DOI: 10.7838/jsebs.2022.27.4.027

[6] C. Wan et al., "Photovoltaic and solar power forecasting for smart grid energy management," CSEE Journal of Power and Energy Systems, Vol. 1, No. 4, pp. 38-46, December 2015. DOI: 10.17775/CSEEJPES.2015.00046

[7] J. Park, J. Moon, S. Jung, and E. Hwang, "Multistep-ahead solar radiation forecasting scheme based on the light gradient boosting machine: A case study of Jeju Island," Remote Sensing, Vol. 12, No. 14, p. 2271, July 2020. DOI: 10.3390/rs12142271

[8] D. So, J. Oh, S. Leem, H. Ha, and J. Moon, "A hybrid ensemble model for solar irradiance forecasting: Advancing digital models for smart island realization," Electronics, Vol. 12, No. 12, p. 2607, June 2023. DOI: 10.3390/electronics12122607

[9] C. Voyant et al., "Machine learning methods for solar radiation forecasting: A review," Renewable Energy, Vol. 105, pp. 569-582, May 2017. DOI: 10.1016/j.renene.2016.12.095

[10] D. So, J. Oh, I. Jeon, J. Moon, M. Lee, and S. Rho, "BiGTA-Net: A hybrid deep learning-based electrical energy forecasting model for building energy management systems," Systems, Vol. 11, No. 9, p. 456, September 2023. DOI: 10.3390/systems11090456

[11] J. Lee, W. Wang, F. Harrou, and Y. Sun, "Reliable solar irradiance prediction using ensemble learning-based models: A comparative study," Energy Conversion and Management, Vol. 208, p. 112582, March 2020. DOI: 10.1016/j.enconman.2020.112582

[12] P. Kumari and D. Toshniwal, "Deep learning models for solar irradiance forecasting: A comprehensive review," Journal of Cleaner Production, Vol. 318, p. 128566, October 2021. DOI: 10.1016/j.jclepro.2021.128566

[13] K. Yan et al., "Short-term solar irradiance forecasting based on a hybrid deep learning methodology," Information, Vol. 11, No. 1, p. 32, January 2020. DOI: 10.3390/info11010032

[14] S. Leem, J. Oh, J. Moon, M. Kim, and S. Rho, "Enhancing multistep-ahead bike-sharing demand prediction with a two-stage online learning-based time-series model: insight from Seoul," The Journal of Supercomputing, September 2023. DOI: 10.1007/s11227-023-05593-6

[15] J. Gui, Z. Sun, Y. Wen, D. Tao, and J. Ye, "A review on generative adversarial networks: Algorithms, theory, and applications," IEEE Transactions on Knowledge and Data Engineering, Vol. 35, No. 4, pp. 3313-3332, April 2023. DOI: 10.1109/TKDE.2021.3130191

[16] X. Huang et al., "Time series forecasting for hourly photovoltaic power using conditional generative adversarial network and Bi-LSTM," Energy, Vol. 246, p. 123403, May 2022. DOI: 10.1016/j.energy.2022.123403

[17] Q. Li, D. Zhang, and K. Yan, "A solar irradiance forecasting framework based on the CEE-WGAN-LSTM model," Sensors, Vol. 23, No. 5, p. 2799, March 2023. DOI: 10.3390/s23052799

[18] J. Moon, S. Jung, S. Park, and E. Hwang, "Conditional tabular GAN-based two-stage data generation scheme for short-term load forecasting," IEEE Access, Vol. 8, pp. 205327-205339, November 2020. DOI: 10.1109/ACCESS.2020.3037063

[19] F. Wang et al., "Generative adversarial networks and convolutional neural networks based weather classification model for day ahead short-term photovoltaic power forecasting," Energy Conversion and Management, Vol. 181, pp. 443-462, February 2019. DOI: 10.1016/j.enconman.2018.11.074

[20] M. N. Fekri, A. M. Ghosh, and K. Grolinger, "Generating energy data for machine learning with recurrent generative adversarial networks," Energies, Vol. 13, No. 1, p. 130, December 2019. DOI: 10.3390/en13010130

[21] C. Tian, C. Li, G. Zhang, and Y. Lv, "Data driven parallel prediction of building energy consumption using generative adversarial nets," Energy and Buildings, Vol. 186, pp. 230-243, March 2019. DOI: 10.1016/j.enbuild.2019.01.034

[22] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, “Modeling tabular data using conditional GAN,” in Advances in Neural Information Processing Systems, Vol. 32, Vancouver, BC, Canada, December 2019. DOI: 10.48550/arXiv.1907.00503

[23] S. Kamthe, S. Assefa, and M. Deisenroth, “Copula flows for synthetic data generation,” arXiv preprint, January 2021. DOI: 10.48550/arXiv.2101.00598

[24] M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein Generative Adversarial Networks,” in Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, August 2017. DOI: 10.48550/arXiv.1701.07875

[25] H. Ishfaq, A. Hoogi, and D. Rubin, “TVAE: Deep metric learning approach for variational autoencoder,” in Proceedings of the Workshop at International Conference on Learning Representations, Vol. 32, Vancouver, BC, Canada, April 2018.

[26] J. Moon, Z. Shin, S. Rho, and E. Hwang, “A comparative analysis of tree-based models for day-ahead solar irradiance forecasting,” in Proceedings of the 2021 International Conference on Platform Technology and Service, pp. 1–6. Jeju, Republic of Korea, August 2021. DOI: 10.1109/PlatCon53246.2021.9680748

[27] L. Breiman, “Random forests,” Machine Learning, Vol. 45, pp. 5–32, October 2001. DOI: 10.1023/A:1010933404324

[28] J. Jang, W. Jeong, S. Kim, B. Lee, M. Lee, and J. Moon, “RAID: Robust and interpretable daily peak load forecasting via multiple deep neural networks and Shapley values,” Sustainability, Vol. 15, No. 8, p. 6951, April 2023. DOI: 10.3390/su15086951

[29] V. Borisov et al., “Deep neural networks and tabular data: A survey,” IEEE Transactions on Neural Networks for Learning Systems, pp. 1–21, December 2022. DOI: 10.1109/TNNLS.2022.3229161

[30] J. Moon, Y. Kim, M. Son, and E. Hwang, “Hybrid short-term load forecasting scheme using random forest and multilayer perceptron,” Energies, Vol. 11, No. 12, p. 3283, November 2018. DOI: 10.3390/en11123283

[31] Orange Data Mining, https://orangedatamining.com.

Authors

Jinyeong Oh has been a dedicated undergraduate student in the Department of AI and Big Data at Soonchunhyang University, South Korea, since 2018. He is set to receive his Bachelor of Science degree in February of next year. Mr. Oh is interested in time-series forecasting, deep learning, generative adversarial networks (GANs), and speech recognition.

Jimin Lee has been a dedicated undergraduate student in the Department of AI and Big Data at Soonchunhyang University, South Korea, since 2020. She is set to receive her Bachelor of Science degree in February of next year. Ms. Lee is interested in natural language processing (NLP), recommendation systems, and multi-modal technologies.

Daesungjin Kim, a dedicated third-grade student at Asan Middle School in Asan, South Korea, exhibits a deep passion for software and artificial intelligence, aiming to contribute significantly to these innovative sectors.

Bo-Young Kim, a key information science educator at Asan Middle School, South Korea, since 2002, earned her Bachelor's and Master's degrees in Information Science and Mathematics Education from Hoseo University in 2002 and 2008, respectively. Ms. Kim is deeply committed to cultivating outstanding students through advanced information education and championing sustainable software training.

Jihoon Moon earned his Ph.D. in Electrical Engineering from Korea University in 2021 and worked as a postdoctoral researcher at Chung-Ang University until 2022. He has been an assistant professor in the Department of AI and Big Data at Soonchunhyang University since 2022. Dr. Moon's research encompasses a broad spectrum, including advanced data mining, information extraction, and time-series analysis, with a keen interest in applying these methods to industrial innovations. His focus extends to sustainable solutions, advanced machine learning, deep learning applications, and explainable AI, with a special emphasis on optimizing energy management and forecasting.