Jurnal Kejuruteraan 37(6) 2025: 2995-3007
https://doi.org/10.17576/jkukm-2025-37(6)-34

Large-Scale Solar Energy Generation Forecasting using Graph Neural Network

Nur Athirah Hasman[a*], Nur Badariah Ahmad Mustafa[a] & Muhammad Nazri Chik[b]

[a]*Institute of Sustainable Energy, Department of Electrical Engineering, College of Engineering, Universiti Tenaga Nasional, Malaysia*

[b]*TNB Renewable Sdn. Bhd.*

**Corresponding author: SE23459@student.uniten.edu.my*

Received 10 July 2024, Received in revised form 22 June 2025
Accepted 22 July 2025, Available online 30 September 2025

ABSTRACT

Solar energy is radiant energy produced by the sun, which radiates and emits enormous amounts of energy every day. Many technologies are used today to harness the energy of the sun, including photovoltaic system. It has promising potential for building large-scale solar power plants. Solar energy forecasts have a significant impact on decisions related to the operation and management of energy systems. Accurately predicting renewable energy performance is critical in ensuring grid reliability, resilience, reducing energy market and system risks. Deep learning (DL) recent success in many applications has attracted researchers implementing DL in solar energy forecasting tasks. Graph Neural Network (GNN) is one of the DL methods that perform inference on data described by graphs. The proposed work implemented GNN method to forecast the solar energy generation of Large-Scale Solar (LSS) farm specifically for LSS Sepang plant. In GNN, the variables in the dataset were described by a self-learning graph structure. Four performance metrics were used to assess the model performance, such as root mean squared error (RMSE), mean squared error (MSE), mean absolute error (MAE) and accuracy. GNN model has demonstrated an effective solar energy generation forecasting model with the lowest value of MAE (537.28), MSE (0.07), RMSE (0.26) and highest accuracy (93.42%), that shows the great potential of the GNN method in forecasting the solar energy generation of LSS farm.

Keywords: Deep learning; graph neural network; large-scale solar plant; solar energy generation forecasting; solar energy

INTRODUCTION

One of the most efficient and promising renewable sources of energy is solar energy which offers a wide range of possible uses. Research continues to prove that solar energy is one of the most eco-friendly, cost-effective, and sustainable energy sources available. Adoption of solar power can protect the Earth by reducing pollution, conserving natural resources, and promoting a greener future for all (Hishammudin et al. 2025). Solar energy generation has a significant influence on the development of solar energy. Due to the influence by solar radiation, ambient temperature, and other weather variables, the output of solar energy varies on a daily, seasonal, and random basis (Ahmad et al. 2020). For large-scale PV grid-connected systems, inconsistent weather variables present considerable issues in terms of power quality, stability, and safety (Basit et al. 2020). Radiation, influence of surroundings, the weather, and the location of the PV panel have a significant impact on how well a photovoltaic (PV) panel performance (Jaleel et al. 2020). The influence of surroundings for example, shading from nearby buildings and tall trees can significantly reduce the amount of solar energy a panel receives and lowering its output. Similarly, panels installed near the coast may suffer from salt mist corrosion, which can degrade materials and diminish efficiency over time. Optimising panel placement and conducting regular maintenance is important to mitigate these negative influences and ensure optimal performance.

Electrical operators at power utility companies must balance electrical consumption and generation in order to conserve energy and avoid its wastage from Renewable Energy (RE) sources. Thus, it is important to estimate the capacity of solar energy output from PV facilities during the upcoming days. The RE sector is currently advancing toward integrating solar power facilities with the electrical grid. The variability and unpredictability of RE sources makes the grid unstable, which also makes it the biggest problem for the sector (Nwaigwe et al. 2019). Therefore, it is important to forecast the amount energy that can be produced by the PV plants. Many researches have been conducted by employing Artificial Intelligence (AI) to anticipate power generation from PV plants and RE farms since AI is a vital step in future development.

AI delivers a good performance in many fields such as in robotics and medical which motivates authors to apply AI technique in solar energy forecasting field. In this work, Graph Neural Network (GNN) method is proposed to describe and self-learn the data using graph structure. GNN represent a set of transition input functions and produce output with more accuracy because it stores a state that can represent information from their neighbourhood with arbitrary depth (Zhou et al. 2020). The method applies neural networks (NN) and cluster analysis to forecasts solar energy generation based on various weather conditions. The advantages of GNN are its ability to analyse the complex data patterns from PV plants and RE farms and able to do prediction or forecasting from multiple stations and areas.

BACKGROUND

SOLAR ENERGY GENERATION FORECASTING

Solar energy is an abundant and renewable resource that has emerged as a pivotal player in the global transition toward sustainable and clean energy solutions. Harnessing the radiant power of the sun through PV systems has become increasingly popular. Solar power plants contribute significantly to the energy grid (Maka et al. 2022). However, the inherent intermittency and variability of solar energy production pose challenges to the effective integration of solar power into the conventional energy grid. Variability in solar energy production is affected by many factors such as cloud cover, atmospheric conditions, and seasonal variations have entailed the development of robust forecasting model to optimise the operational planning and management of solar energy systems (Seckin et al. 2020).

Accurate solar energy generation forecasting holds paramount importance in ensuring grid reliability, economic viability, and sustainable energy practices. With the goal of maximising solar energy utilisation, researchers and practitioners have turned to advanced data-driven techniques, particularly leveraging the capabilities of deep learning methodologies (Impram et al. 2020). The ability of AI techniques to find nonlinear relationships, such as machine learning (ML) algorithms and recently DL model has caught researchers' attention (Benti et al. 2023). The bright future of DL is due to its capacity for generalisation and unsupervised feature learning. The capability of DL has been proven in numerous applications, including image identification, classification work, and natural language processing. Over the past few years, there has been a considerable increase in the number of articles offering deep learning-based approaches, and solar energy forecasting is no exception (Sarker et al. 2021).

OUR CONTRIBUTION

The current study on solar energy forecasting, specifically for LSS plants, shows a notable deficiency in the use of GNN technique. Although there has been significant advancement in ML techniques for forecasting solar energy, there is a noticeable lack of research or models that utilised the capabilities of GNN in this field (Rajagukguk et al. 2020). The lack of study in this area is observed when considering the complex spatial and temporal interconnections that are inherent in LSS energy systems.

The complexity of solar energy systems, especially in the context of large plants, presents challenges that traditional forecasting models may struggle to address effectively. Recognising this gap, the proposition is to leverage GNN method to capture and model the spatial and temporal dependencies within the solar energy system. GNN method offer a novel and promising technique that improves the accuracy of solar energy generated from LSS. Exploring this area of research has great potential, as GNN proven to be effective in capturing intricate connections within data organised in graph structures in several fields (Kamadinata et al. 2019). Utilising GNN for LSS energy forecasting is a great method that can improve the accuracy and dependability of forecasting. The detailed research in this study is needed because better accuracy of forecasting can greatly help in making energy production and management more efficient at LSS.

In conclusion, the research gap that has been found emphasises the need for more investigation and utilisation of GNN techniques in the field of solar energy forecasting, particularly for LSS plants. This is not only tackles the lack of information in existing literature but also creates opportunities for innovative approaches that can contribute to the progress of knowledge in RE research. Addressing this disparity is in line with the overarching objective of maximising the effectiveness and sustainability of extensive solar power systems. This paper explores advancements in solar forecasting focusing on application of GNN, a type of DL that improves forecasting accuracy. The research also addresses challenges using data from various weather conditions and technique to improve forecasting accuracy.

METHODOLOGY

GRAPH NEURAL NETWORK (GNN)

This work proposes to thoroughly explore how time and various features influence forecasting with complex data that changes over time. Additionally, this work aimed to create part in the model that can easy to understand the need to forecast using various features. GNN is a good method in understanding complex relationships and patterns between features and Recurrent Neural Networks (RNN) is useful to forecast data that changes over time (Wang et al. 2018; Liao et al. 2022; Sharma et al. 2023; Morris et al. 2022). RNN is specifically designed to handle sequences of time series data. In this work, the proposed solar forecasting method consist of the combination of GNN and RNN since this work involves the complex interactions between different weather features that influence solar energy generation and forecasted over time series data. However, the main network of the proposed forecasting model is GNN model. This approach enables more accurate and comprehensive forecasting across data that is both interconnected and evolves over time.

GRAPH DEFINITION AND LEARNING

GNN model is formulated as $G = (V,E)$ where V is the set of nodes, and E is the set of edges. N is used to denote the number of nodes in a graph. Basically, a graph must consist of at least two sets of nodes and edges (Lamb et al. 2020; Paudel et al. 2023). In the context of forecasting solar energy generation by GNN model, the concept of nodes (V) and edges (E) can be formally explained by relating them to weather parameters and their interconnections. Therefore, in order to serve as the input to the neural network (NN), the graph must be mathematically processed (Wu et al. 2020).

Node neighbourhood is defined by letting $v \in V$ to denote a node and $e = (v, u) \in E$ to denote an edge which draw a line from u to v. The neighbourhood of node v could be defined as $N(v) = \{u \in V \mid (v, u) \in E\}$ (Yaoma et al. 2021). Besides, the adjacency matrix describes mathematical representation of a graph, denoted as A $\in$ $RN{\times}N$. Aij represents the elements in the adjacency matrix. Where $(vi, vj) \in E$ is an edge connection between i and j when $Aij = c > 0$ (Yang et al. 2021).

Based on the earlier definitions, GNN utilised the adjacency matrix as the input data format. The symmetry of the adjacency matrix also signifies the properties of the graph. A symmetric matrix indicates that the corresponding graph is undirected. In contrast, an asymmetrical adjacency matrix represents a unidirectional graph (Tan et al. 2017). Adopting a graph-based approach, the variables in a multivariate time series is treated as nodes in graphs, as GNN is designed to capture spatial dependencies.

It is important to understand that geographical dependency and temporal dependency are different concepts. Geographical dependency examines the relationships between variables according to the location. Temporal dependency analyses how variables interact over time, emphasising patterns and trends by hour. The graph structure defines the association between multiple variables at a given time point that has no relationship with the nodes specific placement and distance. However, the initial multivariate time series data lacks any adjacency matrix or internal relationship, which will be acquired by the NN itself in this study (Ye et al. 2023). The graph learning layer's methodology as shown in Figure 1 described the construction of an undirected graph and establish relationships. Equations (1), (2) and (3) in Figure 1 described as follows:

$$M1 = \tanh\,(Embedding_1(idx)\Theta_1 \tag{1}$$

$$M2 = \tanh\,(Embedding_2(idx)\Theta_2 \tag{2}$$

$$A = ReLU(\tanh(M_1, M^T{}_2)) \tag{3}$$

where embeddings present two different randomly initialised layers that convert a scalar *idx* vector into an embedding matrix. The embedding layer can be viewed as a layer that maps between different dimensions, and its precise definition can be found in the appropriate references of natural language processing (Hershcovich et al. 2019). In (1) to (2), the variable *idx* represents the node index, which is a positive integer corresponding to the node number. Since the embedding layers are trainable, any changes in the hyperparameters that are dependent on the loss function will impact the generation of the adjacency matrix. In simple terms, the model has the potential to be utilised for graph learning. The model parameters of two simple feed-forward NN are denoted as Θ_1 and Θ_2.

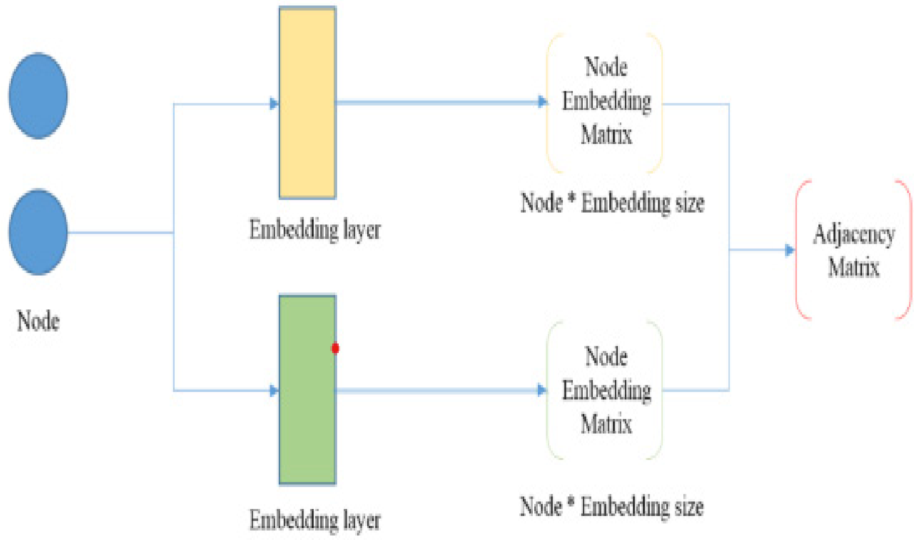


FIGURE 1. The workflow of graph construction in GNN model

In addition, the interpretability of the model can be improved by limiting the maximum number of edges in each node to prevent the model from learning fully connected graphs. The nearest k nodes will be chosen as neighbours for each individual node. A weight of zero is assigned to the nodes that are not connected to maintain the weights for the nodes that are already connected (Wu et al. 2020). The activation function ReLU achieves the symmetric property of the suggested graph adjacency matrix. At this stage, constructing the graph for the numerous time series nodes had finished, and the graph learning process was carried out using gradient descent. Utilising the loss function for parameter updates in the graph learning network enables the network to acquire the optimal graph structure. Once the learning network converges, the optimal graph accurately presents the spatial relationships among various features.

DATA SAMPLING

Various features and labels used in solar energy forecasting usually in time series data, arranged in two-dimensional (2D) data namely time steps and feature columns. However, for long short-term memory (LSTM) and graph convolutional neural networks (GCN), the dimensionality of the input data is often above three-dimensions (3D) which includes batch, time step and feature. LSTM is a type of RNN that is designed to process and forecast outcomes in sequential data which focused on capturing long-term dependencies. GCN focus on analysing data structured between nodes. In simple words, LSTM address the challenges of sequences over time while GCNs focus on the complexities within graph-structured data. Another requirement of the batch gradient descent algorithm is to sort the data into the corresponding format so that the data will smoothly receive and calculate by the model (Goodfellow et al. 2016).

The conversion of 2D data to 3D data is done by adding a batch dimension aligns with the requirements of LSTM and GCN models. This transformation enables efficient batch processing with respect to the sequential nature of time series data, and accommodates multiple features per time step, facilitating more sophisticated and accurate modelling of complex temporal dynamics (Goodfellow et al. 2016). As shown in Figure 2, data sampling process consist of two processing methods described as expansion in the time scale and expansion in different variables. These processing methods change the data into format that GNN model can understand. This allows the model to use all information in the data like how different parameters are related over time to make better and detailed forecasting.

TIME SCALE AND VARIABLE EXPANSION

Expanding over time is appropriate for the input format of the LSTM network as it often needs a three-dimensional input tensor with dimensions (batch, time step, feature). In order to align the data format with the input structure of the LSTM network, the sliding window technique is employed to partition and arrange the input data (Mishra et al. 2023).

The variables, X_1 to X_2 in Figure 2 denoted by solid lines of various colours, reflecting the distinct components of the multivariate time series. In order to fulfil the data requirement of the LSTM network which include batch, time step and feature, the sliding window technique was employed to extract data samples (Kapoor et al. 2013), as shown by the moving window in Figure 2. The moving step denotes the magnitude of each movement made by the moving window. In this work, the step size of the moving window is fixed at 1 due to limited data samples and to enhancing data utilisation. This approach makes every possible sequence within the dataset is used, enhancing the model's performance to learn subtle patterns and shifts that probably are missed with larger step sizes. Besides, it achieves a balance between enhancing the dataset to improve forecasting accuracy and handling computational complexity, making it a realistic option for thorough and effective time series analysis.

The expansion in the time scale was executed based on the blue dashed box shown in Figure 2. Expansion in time dimension described the difference in data at different time steps. After using the moving window to obtain fixed-length of time-series data, the features of each time step comprise all the variables in the time step. Features represents as spatial dependencies such as temperature or humidity in a dataset that are input into a model. The data of a single moving window could be expressed as a two-dimensional array of time step and variable numbers. Additionally, as the number of moving windows set to 1 step size increases, the data could be expanded easily into 3D data through the random selection of different moving windows. This feature makes it easier to perform batch gradient descent in the final DL model.

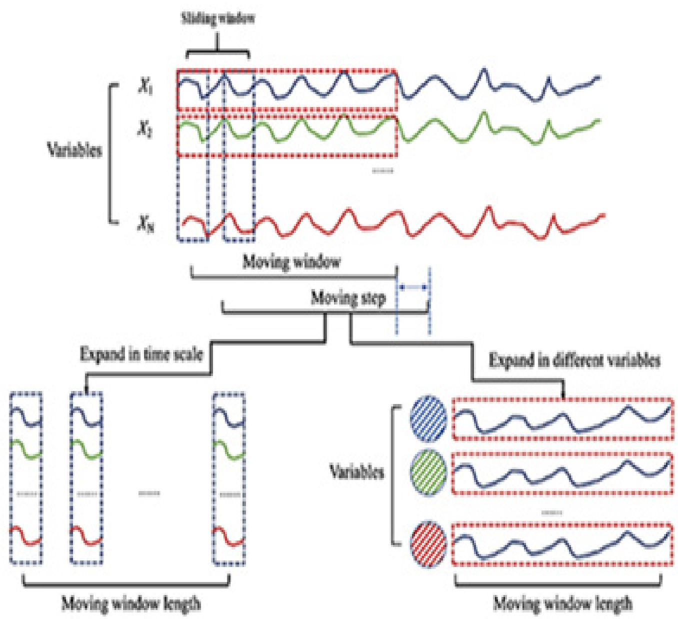


FIGURE 2. Data sampling process

GNNs are typically not used for analysing time series data. Therefore, it is crucial to have an effective approach for processing raw data. The graph network treats variables as nodes. Once the graph structure has been acquired, the determination of the feature definition for each node becomes crucial for the computation and training of the GNN. It presumes that the characteristics of each node corresponded to the time-series data of each variable. Using all the time-series data as features for the nodes is not recommended due to the significant computing burden it imposes. As a result, a data imbalance issue may arise when the model's spatial and temporal dependencies are not properly accounted for, leading to difficulties in training and achieving convergence (Li et al. 2024).

PERFORMANCE EVALUATION METHOD

The evaluation variables utilised to assess the forecasting accuracy of each model for solar energy were root mean squared error (RMSE), mean squared error (MSE), mean absolute error (MAE) and accuracy (Granderson et al.

2015). The evaluation metrics demonstrate a variance in accuracy between the forecasts and historical data, which serves as an indication of the model's capability to forecast the total solar energy in the test dataset. However, considering the small value of absolute value after normalised the solar energy reading may also affect the error values of MAE, MSE, and RMSE that will be exceedingly small. Therefore, small modifications to the calculation procedure of MAE, MSE, and RMSE has been made as well as adjusted the averaging timeframe to one day. The equation of MAE, MSE, RMSE and accuracy are shown in (4) to (7), respectively.

$$MAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{N} \tag{4}$$

$$MSE = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{N} \tag{5}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{N}} \tag{6}$$

$$Accuracy = \frac{Total\ number\ close\ matches}{Total\ number\ of\ samples} 100\% \tag{7}$$

where y, $\hat{y}$, n, and N in illustrate the measured data, predicted data, number of test datasets, and number of predicting days, respectively.

LSS SOLAR ENERGY GENERATION DATASET

The weather dataset obtained from the Sepang solar farm consist of one year hourly measured meteorological data in Sepang from October 21, 2020 to October 20, 2021. The dataset includes 14 features; numeric features (10) and categorical features (4). The numeric features were atmospheric pressure (Sensor 1 and 2) (hPa), average ambient temperature (C), horizontal irradiance (W/m2), humidity, panel temperature (C), tilted irradiance (W/m2), wind direction, rain, and wind speed (m/s). While, another four features were measured hourly, obtained from the measuring equipment placed at the solar plant such as day of the year, week of the year, month, and hour. All categorical features and numeric features with various reading shown in Figure 3.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	time	Atmospheri	Atmosphe	Average A	Horizonta	Panel Ten	Rain (Sens	Humidity	Tilted Irra	Wind Dire	Wind Spe	day_of_ye	week_of_	month	hour	AC Energy Total (kWh)
2	21/10/2020 0:00	1009.9	1011.2	23.6	0	24.6	0	86.1	0	130.8	0.4	295	43	10	0	0.5
3	21/10/2020 1:00	1009.9	1011.2	23.6	0	24.6	0	86.1	0	130.8	0.4	295	43	10	1	0
4	21/10/2020 2:00	1010.3	1011.6	23.8	0	25.1	0	85.1	0	51.6	0.9	295	43	10	2	0
5	21/10/2020 3:00	1010	1011.3	23.9	0	25	0	84	0	91.8	0.8	295	43	10	3	0
6	21/10/2020 4:00	1009.5	1010.8	23.7	0	25	0	81.6	-6.7	41.6	1.6	295	43	10	4	0
7	21/10/2020 5:00	1009	1010.3	23.5	0	24.9	0	85.3	0	62.7	1	295	43	10	5	0
8	21/10/2020 6:00	1008.9	1010.2	23.1	0	25.2	0	85.6	0	38.7	1.3	295	43	10	6	0
9	21/10/2020 7:00	1008.9	1010.2	23.1	0.2	24.5	0	86.2	0.2	248	0.5	295	43	10	7	0
10	21/10/2020 8:00	1009.7	1011	23.3	71.3	25.5	0	86.1	73.3	67	1.2	295	43	10	8	0.054216867
11	21/10/2020 9:00	1010.3	1011.6	24	391.4	31.9	0	78	391.7	91.9	1.1	295	43	10	9	0.325301205
12	21/10/2020 10:00	1010.6	1011.9	25.4	649.3	48.6	0	65.4	642.6	247.5	1.2	295	43	10	10	0.518072289
13	21/10/2020 11:00	1010.9	1012.2	25.7	805.1	51.1	0	55.2	802.4	281	1.9	295	43	10	11	0.638554217
14	21/10/2020 12:00	1010.5	1011.8	26.4	308.7	47.5	0.1	57.1	305	256.4	2	296	43	10	12	0.234939759
15	21/10/2020 13:00	1009.8	1011.1	26.1	434.8	40.1	0	71.7	432	187.5	2.3	296	43	10	13	0.34939759
16	21/10/2020 14:00	1008.9	1010.2	24.3	177	26.9	-0.7	82.2	175.2	267	3.3	296	43	10	14	0.15060241
17	21/10/2020 15:00	1007.7	1009	24.2	78.9	25.4	0	87.2	79.2	229.8	1.6	296	43	10	15	0.060240964
18	21/10/2020 16:00	1007.1	1008.4	23.9	71.2	26.6	0	84.5	70.9	292.2	1.2	296	43	10	16	0.060240964
19	21/10/2020 17:00	1006.8	1008.1	24.1	76.1	26.9	0	84	74.7	316.4	1.2	296	43	10	17	0.06626506
20	21/10/2020 18:00	1007.2	1008.5	24.6	65.3	28.4	0	83	64.5	48.1	1.4	296	43	10	18	0.048192771
21	21/10/2020 19:00	1007.5	1008.8	24.4	10.3	27	0	84.2	10.6	57.6	1.1	296	43	10	19	0.006024096
22	21/10/2020 20:00	1007.9	1009.2	24.3	0	25.5	0	85.6	0	76.4	0.5	296	43	10	20	0
23	21/10/2020 21:00	1008.8	1010.1	24.2	0	25.4	0	86	0	143.9	0.8	296	43	10	21	0
24	21/10/2020 22:00	1010.2	1011.5	24.1	0	25.5	0	87.5	0	99.4	0.6	296	43	10	22	0
25	21/10/2020 23:00	1010.5	1011.8	24.3	0	25.8	0	87.2	0	200.6	0.4	296	43	10	23	0
26	22/10/2020 0:00	1010.5	1011.8	24.1	0	25.3	0	87.8	0	176.5	0.7	296	43	10	0	0
27	22/10/2020 1:00	1010.1	1011.4	24.1	0	25.3	0	88	0	133.8	0.5	296	43	10	1	0

FIGURE 3. Screenshot of data arrangement

The dataset in this work was organised based on the time period, and all models utilised the identical division process. The training dataset for the model consisted of data collected between October 21, 2020 and May 20, 2021. The data collected between May 21, 2021 to October 20, 2021 was utilised as a validation set to fine-tune which is the most effective model parameters. The test set comprised 153 continuous days of data, running from May 21, 2021 to October 20, 2021. However, due to the network requirement of encoding the data from the previous 24 hours, the forecasting results began on October 21, 2021 and covered a period of 152 days and 3648 hours. In order to optimise the computational efficiency of data processing, the dimensions of the sliding window were set to a fixed value of 24. As a result, the daily frequency will be used within the data structure. Other extended distance periodic networks could benefit from incorporating the input characteristics such as day of year, week of year, and month.

DEFINITION OF GNN MODEL

The evaluation of forecasting performance was based on the iterative process of development of models and subsequent validation, with a focus on reaching better levels of accuracy. It is established that the models in this investigation generated one-hour predictions. The forecasting model utilise the GNN methodology to predict the solar energy generation of LSS systems. Figure 4 depicts the structure of the GNN computational model. This model is built using eight specifics weather parameters: air pressure (measured by sensor 1 and 2), average ambient temperature, horizontal irradiance, panel temperature, tilted irradiance, wind direction, and wind speed.

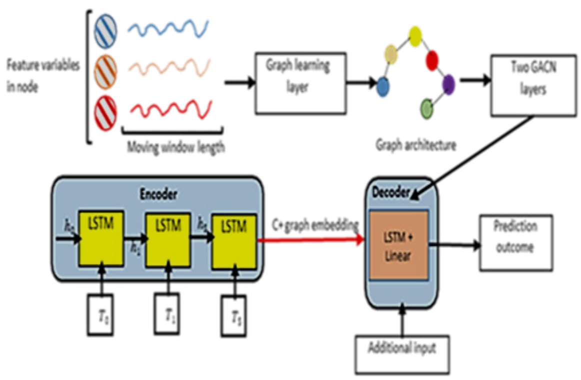


FIGURE 4. GNN architecture computational module

The GNN model utilises data samples spanning from 00:00 am to 23:00 pm to forecast the total energy generation for the following day. The chosen time frame is based on the gradual rise in horizontal irradiance values, indicating the ideal duration for solar panels to gather irradiance data. After this specific period, solar panels generally do not gather substantial amounts of irradiance. The list of input parameters for the forecasting system does not include samples of humidity and rain data. This judgement is made because humidity and rainfall show the negative correlation with total energy as determined using correlation matrix analysis and statistical analysis.

Graph Attention Convolutional Network (GACN) and LSTM were implemented to capture both spatial and temporal relationships concurrently. The graph structure acquired by the model was regarded as an indication of spatial interconnections. The NN analyses the graph structure to understand the spatial relationship between various external variables and the goal variable in prediction. In order to enhance the comprehensibility of the graph model, it is advisable to restrict the maximum number of edges in the graph structure to 45.

GNN model algorithms were written in Python 3.7 and Google Colab. Table 1 shows the structural parameters and addition hyperparameters of GNN model development. The parameters value for the GNN model were selected based on the previous studies and practical testing to ensure a balance between effective performance and efficient computational modules. These choices are detailed and supported by references to relevant research, ensuring the model's approach by proven techniques (Gao et al. 2022).

TABLE 1. Model Structural Parameters

Structures of network	Other parameters	Features
Encoder: Graph learning layer (node features: 40), GACN (1 layer, heads = 8), LSTM (hidden units 128); Decoder: LSTM (hidden units 128), 2 Linear layers (units 256)	Activation function: Leaky Relu; Optimizer: Adam Dropout = 0.1	Encoder length: 24; Encoder features: 13; Node number: 13

RESULTS AND DISCUSSION

OVERVIEW FORECAST RESULT

After the model was develop, the performance of the model was evaluated by measuring MAE, MSE, RMSE, and accuracy. MAE is a significant metric that quantifies the size of prediction mistakes by calculating the average absolute errors in forecasts. It acts as a critical indicator of the correctness of the model. A lower MAE indicates higher precision in predictions, demonstrating the model's capacity to closely align with the actual data (Petropoulos et al. 2022).

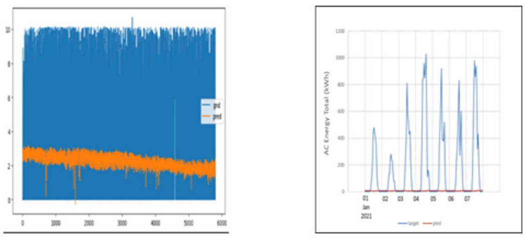
Other performance metrics MSE and RMSE provide information about the variability and standard deviation of mistakes. The MSE highlights significant errors more prominently by squaring the differences. Consequently, a lower MSE value signifies enhanced accuracy. RMSE is a statistical measure that quantifies the standard deviation. It enhances the evaluation process by providing a metric that uses the same unit as the actual values being measured. A lower RMSE indicates higher accuracy of the model that also describes the model capability of making predictions that closely align with the distribution of the observed data. Additionally, the model can be evaluated by measuring its accuracy, which quantifies the proportion of right classifications and provides a comprehensive assessment of the model's performance. However, it is important to exercise caution when dealing with imbalanced datasets, as a high accuracy might be deceptive in such cases.

In the context of forecasting solar energy using GNN, determining whether the dataset is imbalanced involves assessing the distribution of different weather conditions, time periods, or any other categorisation relevant to energy output. The primary criterion for identifying imbalance is a significant disproportion in the number of samples across these categories or classes. The number of samples in each category is calculated and compared and the distribution of continuous forecast values should be examined for skewness to identify imbalances. The development of the GNN model for forecasting solar energy generation having an improvements and updates phase by phase. Starting from phase 1 in outcome 1, where the result of accuracy in forecasting solar energy was very low until successfully reached higher accuracy which will be explained in outcome 5. A step-by-step enhancement of the solar energy prediction model was conducted and explained in each outcome (outcome 1 to 5). Initially in outcome 1, the model's performance was assessed using metrics such as errors and accuracy, although no improvements were made at this stage. In outcome 2, data faced the rearrangement process and normalised process.

Subsequently in outcome 3, a reorganisation of temporal data, including days and months, was carried out to better capture patterns, leading to a slight increase in prediction accuracy. In the following outcome 4, a selection of the most influenced features for forecasting purposes was done, which resulted in a notable improvement in the model's accuracy. This was followed by an optimisation of the data splitting process in outcome 5, adopting a 60%-40% split for training and testing, respectively, thereby facilitating more effective learning and testing outcomes. Finally, the GNN model validate using another data from different solar farm which is Gambang Solar Farm. Through these phases, carried out in a structured and methodical manner, the model was progressively refined from its initial assessment to achieving excellent performance in practical applications.

OUTCOME 1: USING LSS SEPANG RAW DATA

The aim of this outcome is to create a preliminary GNN model for forecasting solar energy generation. In this initial stage, the model demonstrated a MAE of 1698210.93. The MSE and RMSE were 197933.11 and 444.90, respectively. The initial accuracy was recorded at a negligible 0.02%. Despite subsequent evaluations, there was no noticeable gain in accuracy, as evidenced by the progress percentage remaining stagnant at 0%. Figure 5 shows the forecasting outcome between prediction data (orange) and the raw data (blue). As shown in Figure 5, the prediction value was far lower compared to raw data. From the findings, it can be concluded that in outcome 1, the model's performance has not progressed beyond the initial assessment, with mistakes remaining at their original levels and accuracy failing to enhance. Hence, it is recommended to improve the forecasting skills by rearranged the data and normalised data to make sure the dataset fit the model perfectly in the next outcome.



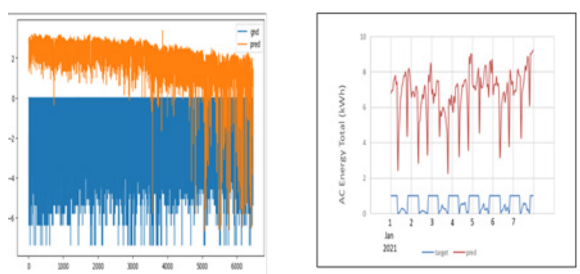
(a) Prediction (orange) vs historical data (blue)

(b) Prediction (red) vs historical data (blue) in one week (1-7 January 2021)

FIGURE 5. Outcome 1: solar energy generation. (a) Generated using Python, (b) Prediction for one week in January 2021

OUTCOME 2: DATA NORMALISATION

The second outcome aims to combine diverse data sources and boost the accuracy of the model by improving the quality of the data through data normalisation. Figure 6 shows the forecasting outcome between prediction data (orange) and actual data (blue). As depicted in Figure 6, the prediction value is still far lower compared to actual data. Figure 6(b) also shows the prediction value (red colour) and actual value is quite far even after normalized the dataset. The outcome 2 assessment yielded a MAE of 27745.78. Additionally, the evaluation resulted in a MSE of 20.86 and RMSE of 4.57. The original accuracy was measured to be 0.02%. After the data was normalized, there was a significant improvement in accuracy value to 1.16%, indicating a progress rate of 1.14%.



(a) Prediction (orange) vs historical data (blue)

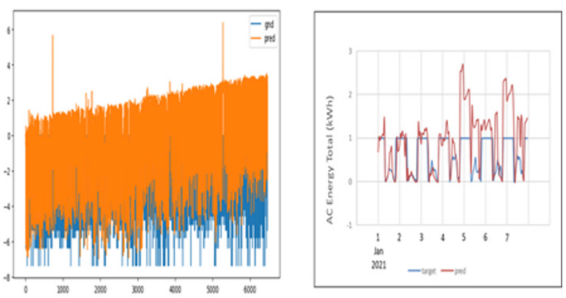
(b) Prediction (red) vs historical data (blue) in one week (1-7 January 2021)

FIGURE 6. Outcome 2: solar energy generation. (a) Generated using Python, (b) Prediction for one week in January 2021

Prior to outcome 2, the model encountered significant difficulties, which were marked by considerable mistakes such as wrongly handling the dataset according the GNN needed, as demonstrated by the higher MSE and RMSE values. The measurements revealed substantial discrepancies between the predicted and actual results, making the model untrustworthy. Due to that, the implementation of strategic measures, such as integrating extensive data sources and standardising the dataset, led to a significant improvement. The MSE decreased to 20.86, showing a significant decrease in errors. Additionally, the RMSE decreased to 4.57 from 444.90, suggesting a large improvement in the consistency of predictions. These enhancements represent a significant improvement in the accuracy and dependability of the model, indicating a great progress towards attaining the project's goals.

OUTCOME 3: ARRANGEMENT OF CATEGORICAL FEATURES

The third outcome involves incorporating time characteristics, such as day of year, week of year, month, and hour, as categorical features for each data point. This ensures that particular values are assigned to each data point. Figure 7 shows the forecasting outcome between prediction data (orange) and actual data (blue). As seen in Figure 7, the prediction value showing good improvement where the prediction outcome is nearer towards actual data. Figure 7(b) also depicted the better prediction value outcome illustrated in one week where MAE, MSE and RMSE were 17845.89, 15.78 and 3.97 respectively.



(a) Prediction (orange) vs historical data (blue)

(b) Prediction (red) vs historical data (blue) in one week (1-7 January 2021)

FIGURE 7. Outcome 3: solar energy generation. (a) Generated using Python, (b) Prediction for one week in January 2021

In outcome 3, a crucial approach was employed that entailed careful organisation of categorical attributes that depict temporal qualities such as day of the year, week of the year, month, and hour. The features were designed to assign precise and important values to each data point, enabling the model to effectively capture temporal trends and greatly improve its forecast accuracy. The significant decrease in MAE metrics from 27745.78 to 17845.89 demonstrates a substantial enhancement in prediction precision. The reduction in MSE metrics from 20.86 to 15.78 and RMSE metrics from 4.57 to 3.97 also suggests the improved model accuracy and consistency in predictions. The most notable accomplishment is the huge improvement in accuracy, which rose from an initial 1.16% to 24.84% after integrating categorical characteristics related to time. This highlights the effectiveness of this strategy in greatly enhancing the model's accuracy.

OUTCOME 4: PARAMETER OPTIMISATION BY CORRELATION ANALYSIS

In this section, the improvement done involved optimising model parameters by correlation analysis. This study was used to pick the most influential elements and enhance the accuracy of predictions. Figure 8 shows the correlation analysis between weather parameters with solar total generated energy. Positive correlations indicate that as the values of the parameter increase, the total generated energy tends to increase as well. The strongest positive correlations were horizontal irradiance (W/m2), tilted irradiance (W/m2), and panel temperature (C). On the other hand, humidity (%) shows a strong negative correlation. As the humidity increases, the AC Energy Total tends to decrease. This suggests that high humidity may have a negative impact on energy generation. The parameters atmospheric pressure (sensor 1 and 2) (hPa), average ambient temperature (C), and rain (Sensor 1) show no significant correlation with the total generated energy.

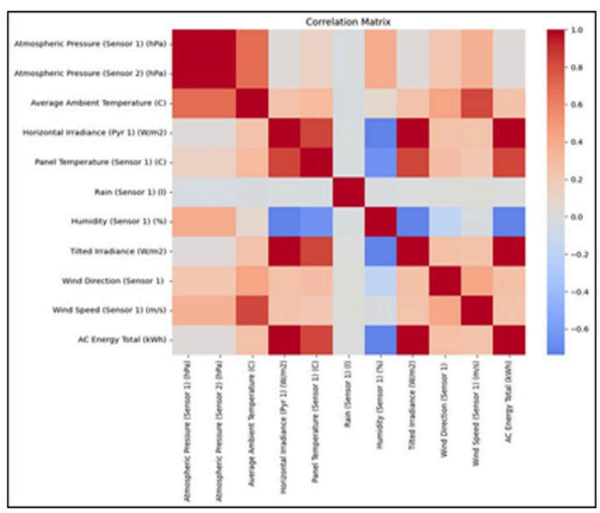
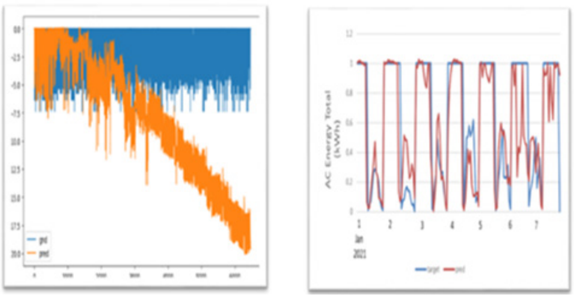


FIGURE 8. Correlation Analysis

Figure 9 shows the forecasting outcome between prediction data (orange) and actual data (blue). The prediction value become closer towards actual data compared to Figure 7. The better prediction outcome can be seen in Figure 8(b) where MAE, MSE and RMSE decreased to 3560.96, 0.44 and 0.67 respectively. The decreases demonstrate a substantial enhancement in forecasting precision and consistency. The model accuracy was improved from 24.84% to 48.38% that depicted effectiveness of picking the most relevant parameters in GNN model.



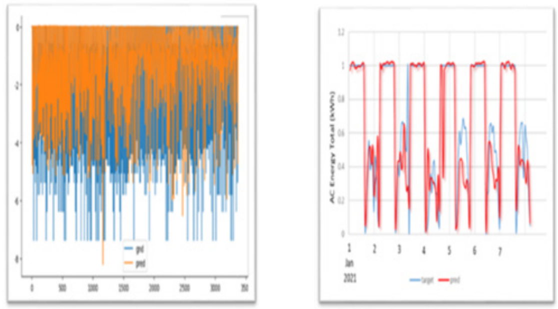
(a) Prediction (orange) vs historical data (blue)

(b) Prediction (red) vs historical data (blue) in one week (1-7 January 2021)

FIGURE 9. Outcome 4: solar energy generation. (a) Generated using Python, (b) Prediction for one week in January 2021

OUTCOME 5: TRAINING AND TESTING SPLIT OPTIMISATION (60%-40%)

In outcome 5 taking an action on implementing a 60%-40% split between training and testing datasets to enhance forecasting accuracy through optimised data allocation. Figure 10 shows the forecasting outcome between prediction data (orange) and actual data (blue). As seen in Figure 10, the prediction value shown significant improvement compared to Figure 9. Figure 10(b) also depicted a better performance between prediction value (red) and actual value (blue), where both were very close towards each value. In outcome 5, the GNN model exhibited better performance described by MAE, MSE and RMSE, 537.28, 0.07 and 0.26 respectively, suggesting enhanced forecasting precision and consistency. The model accuracy was increased from 48.38% to 93.42% highlights the effectiveness of the optimised division of training and testing, demonstrating a stronger and more precise model.



(a) Prediction (orange) vs historical data (blue)

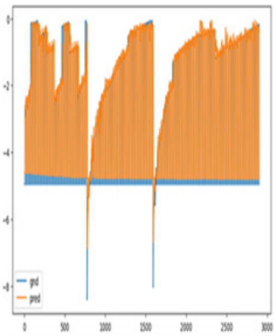
(b) Prediction (red) vs historical data (blue) in one week (1-7 January 2021)

FIGURE 10. Outcome 5: solar energy generation. (a) Generated using Python, (b) Prediction for one week in January 2021

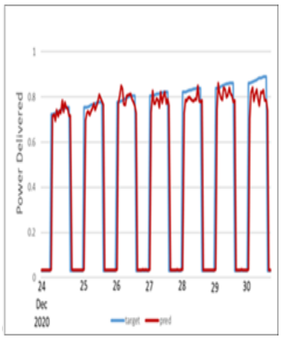
This outcome emphasised on optimising the division of training and testing data, specifically by adopting a ratio of 60% (training data) and 40% (testing data). The deliberate distribution of data enabled a more efficient training of the model, resulting in improved predicted precision. The optimised split ensured that the model was trained on a significant proportion of the data, allowing it to capture complex patterns within the dataset. The thorough approach made a substantial contribution to the enhanced accuracy of the model.

VALIDATION USING LARGE-SCALE GAMBANG DATASET

The model underwent testing with Gambang data to verify its accuracy in real-world scenarios. The objective was to examine the model's accuracy in forecasting solar energy generation using data from other solar farms. The accuracy and reliability of the GNN model were validated using Gambang data. This helps to confirm that the model works well in real-world situations and effectiveness. The process involved in achieving this outcome entails meticulous changes and refinements to the model in order to ensure its predictions were in line with the distinct patterns observed in Gambang data. Figure 11 shows the forecasting outcome between prediction data (orange) and actual data (blue). As seen in Figure 11(a), the prediction value showing better improvement where the prediction outcome was closer to actual data. This finding was proven in Figure 11(b). Using Gambang data, the model performance depicted lower MAE, MSE and RMSE; 99.19, 0.01 and 0.09, respectively. The model was tuned a bit to match with the Gambang datasets characteristic. The lower MAE, MSE, and RMSE values indicates a good forecasting result and higher accuracy value, 98%. The findings reflect that the GNN model is capable to forecast solar energy generation with precision, even having complex and varied data patterns.



(a) Prediction (orange) vs historical data (blue)



(b) Prediction (red) vs historical data (blue) in one week (24 - 30 December 2020)

FIGURE 11. Validation of the model. (a) Generated using Python, (b) Prediction for one week in December 2020

CONCLUSION

The model underwent a thorough transformation throughout various stages from outcome 1 until outcome 5 by improving modifications with the aim of improving the GNN model ability to make accurate forecasting. In outcome 1, the initial progress showed the model needs more improvement, since the model resulted higher errors of MAE (1698210.93), MSE (197933.11), RMSE (444.90) and low accuracy (0.02%). Outcome 2 witnessed significant improvement which enhance the model's accuracy through improvised the data quality and normalising the data set. In outcome 3, another improvement of accuracy was observed due to implement the arrangement of categorical features (day of year, week of year, month, and hour) as time characteristics for each data point. This change resulted a big improvement in both accuracy and low errors on GNN model.

Outcome 4 of the research aimed to further enhance the model accuracy by optimising model parameters using correlation analysis to select the most influential features. This outcome resulted higher accuracy compared to the previous outcome. In outcome 5, there was a work that focus on improving the way of the training and testing data. The model implements a 60%-40% split between training and testing datasets to enhance predictive accuracy through optimised data allocation. This technique resulted the highest accuracy which conclude that the model can managed complex patterns in the dataset. The model resulted higher errors of MAE (537.28), MSE (0.07), RMSE (0.26) and low accuracy (93.42%). The comparison of outcome 1 and outcome 5 proved that the GNN model developed well. Lastly, the efficiency of the model can be observed through the validation of GNN model using another dataset from other solar farm. LSS Gambang was chosen to validate model and resulting an outstanding accuracy with MAE (99.19), MSE (0.01), RMSE (0.09) and accuracy (98%).

In summary, significant improvements in accuracy were achieved through an iterative progress that involved the improving data quality, the arrangement of categorical features, the optimisation of parameters, and training and testing split optimisation of data. The ability of the model to accurately forecast solar energy generation even when dealing with a variety of datasets will maintain the robustness and effectiveness in real-world scenarios.

ACKNOWLEDGEMENT

The authors would like to thank Ministry of Higher Education under the TNB Seed Fund number: [U-TV-RD-21-18] for funding this project.

DECLARATION OF COMPETING INTEREST

None.

REFERENCES

Ahamad Sabki, S., Hishamuddin, H., Sabtu, M. I., & Wahid, Z. 2025. Aplikasi tenaga solar dalam industri pembuatan Malaysia: Analisis cabaran dan faktor kejayaan. *Jurnal Kejuruteraan* 37(1).

Ahmed, R., Sreeram, V., Mishra, Y., & Arif, M. D. 2020. A review and evaluation of the state-of-the-art in PV solar power forecasting: Techniques and Optimization. *Renewable and Sustainable Energy Reviews* 124: 109792.

Basit, M. A., Dilshad, S., Badar, R., & Sami ur Rehman, S. M. 2020. Limitations, challenges, and solution approaches in grid-connected renewable energy systems. *International Journal of Energy Research* 44(6): 4132–4162.

Benti, N. E., Chaka, M. D., & Semie, A. G. 2023. Forecasting renewable energy generation with machine learning and Deep learning: Current advances and future prospects. *Sustainability* 15(9): 7087.

Goodfellow, I., Bengio, Y., & Courville, A. 2016. *Deep Learning.* MIT Press.

Granderson, J., Touzani, S., Custodio, C., Sohn, M., Fernandes, S., & Jump, D. 2015. Assessment of Automated Measurement and Verification (M&V) Methods.

Gao, Y., Miyata, S., & Akashi, Y. 2022. Interpretable deep learning models for hourly solar radiation prediction based on graph neural network and attention. *Applied Energy* 321: 119288.

Hershcovich, D., Toledo, A., Halfon, A., & Slonim, N. 2019. Syntactic interchangeability in word embedding models. *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations*.

Impram, S., Varbak Nese, S., & Oral, B. 2020. Challenges of renewable energy penetration on power system flexibility: A survey. *Energy Strategy Reviews* 31: 100539.

Jaleel, Emad & Resen, Ali & Nmr, Jawad & Mahdi, Emad. 2020. The behavior of the solar panel temperature under different environmental conditions. *Test Engineering and Management* 83: 7235 - 7243.

Kamadinata, J. O., Ken, T. L., & Suwa, T. 2019. Solar irradiance fluctuation prediction methodology using artificial neural networks. *Journal of Solar Energy Engineering* 142(3).

Kapoor, P., & Bedi, S. S. 2013. Weather forecasting using sliding window algorithm. International Scholarly Research Notices.

Liao, W., Bak-Jensen, B., Radhakrishna Pillai, J., Wang, Y., & Wang, Y. 2022. A review of graph neural networks and their applications in power systems. *Journal of Modern Power Systems and Clean Energy* 10(2): 345–360.

Li, Z., Gao, Z., Zhang, G., Liu, J., & Xu, L. 2024. Dynamic personalized graph neural network with linear complexity for multivariate time series forecasting. *Engineering Application of Artificial Intelligence* 127: 107291.

Lamb, L. C., Garcez, A. d'Avila, Gori, M., Prates, M. O. R., Avelar, P. H. C., & Vardi, M. Y. 2020. Graph Neural Networks meet neural-symbolic computing: *A survey and perspective. Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*.

Maka, A. O., & Alabid, J. M. 2022. Solar energy technology and its roles in sustainable development. *Clean Energy* 6(3): 476–483.

Morris, C. 2022. Graph Neural Networks: Graph Classification. Graph Neural Networks: Foundations, Frontiers, and Applications, 179–193.

Mishra, P. 2023. Explainability for deep learning models. *Explainable AI Recipes*, 233–250.

Nwaigwe, K. N., Mutabilwa, P., & Dintwa, E. 2019. An overview of solar power (PV systems) integration into electricity grids. *Materials Science for Energy Technologies* 2(3): 629–633.

Paudel, D., de Wit, A., Boogaard, H., Marcos, D., Osinga, S., & Athanasiadis, I. N. 2023. Interpretability of deep learning models for crop yield forecasting. *Computers and Electronics in Agriculture* 206: 107663.

Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M. Z., Barrow, D. K., Ben Taieb, S., Bergmeir, C., Bessa, R. J., Bijak, J., Boylan, J. E., Browell, J., Carnevale, C., Castle, J. L., Cirillo, P., Clements, M. P., Cordeiro, C., Cyrino Oliveira, F. L., De Baets, S., Dokumentov, A., … Ziel, F. 2022. Forecasting: Theory and practice. *International Journal of Forecasting* 38(3): 705–871.

Rajagukguk, R. A., Ramadhan, R. A., & Lee, H.-J. 2020. A review on Deep Learning Models for forecasting time series data of solar irradiance and photovoltaic power. *Energies* 13(24): 6623.

Seckin Salvarli, M., & Salvarli, H. 2020. For sustainable development: Future trends in renewable energy and Enabling Technologies. *Renewable Energy - Resources, Challenges and Applications*.

Sarker, I.H. 2021. Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. *SN COMPUT. SCI.* 2: 420.

Sharma, A., Sharma, A., Nikashina, P., Gavrilenko, V., Tselykh, A., Bozhenyuk, A., Masud, M., & Meshref, H. 2023. A Graph Neural Network (gnn)-based approach for real-time estimation of traffic speed in Sustainable Smart Cities. *Sustainability* 15(15): 11893.

Tan, Z., Zhao, X., & Wang, W. 2017. Representation learning of large-scale knowledge graphs via entity feature combinations. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*.

Wang, F., Yu, Y., Zhang, Z., Li, J., Zhen, Z., & Li, K. 2018. Wavelet decomposition and convolutional LSTM networks based improved deep learning model for solar irradiance forecasting. *Applied Sciences* 8(8): 1286.

Wu, Z., Pan, S., Long, G., Jiang, J., Chang, X., & Zhang, C. 2020. Connecting the dots: Multivariate time series forecasting with Graph Neural Networks. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

Yaoma & Jiliang. 2021. Scalable graph neural networks. *Deep Learning on Graphs*, 162–175.

Yang, C., Shi, C., Liu, Z., Tu, C., & Sun, M. 2021. Network embedding for large-scale graphs. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 99–117.

Ye, J., Liu, Z., Li, W., Zhu, T., Sun, L., & Du, B. 2023. Evolving Graph Structure Learning for Multivariate Time Series Forecasting.

Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., & Sun, M. 2020. Graph Neural Networks: A review of methods and applications. *AI Open* 1: 57–81.