

Disaggregating Distributed PV Power from Aggregate Measurements in Transmission Systems

Hediyeh Safari, *Student Member, IEEE*, Claudio A. Cañizares, *Member, IEEE*, Daniel Sohm, *Member, IEEE*, Elyas Ahmed, *Member, IEEE*, Ismael El-Samahy, *Member, IEEE*, and Maurice Dusseault

Abstract—The growing integration of Photovoltaic (PV) systems into distribution networks has limited visibility for system operators, as the power output of low-power PV systems is not typically monitored. The study presented in this paper introduces a novel approach for disaggregating PV generation from net-load measurements recorded at the transmission system level. The proposed technique is applied to actual data from an existing transmission line feeder to identify PV generation from net demand measurements. The developed methodology incorporates Geographical Information Systems (GIS) for detecting PV panels based on an algorithm that provides an accurate estimate of solar PV capacity. These estimations are validated against actual data from a local utility, showing a close match between the two. This information is then utilized in reliable software tools to simulate PV power generation in the studied region, which is then used to estimate and disaggregate the generated power from net-load data by applying multiple Machine Learning and Deep Learning models. The results demonstrate that, with the proposed approach, it is feasible to adequately disaggregate PV power generation from transmission feeder net-load measurements with minimal or no additional sensor infrastructure.

Index Terms—Data disaggregation, data-driven analysis, distributed solar power, Geographical Information Systems (GIS), machine learning, transmission systems.

NOMENCLATURE

Acronyms

BiLSTM	Bidirectional LSTM
CNN	Convolutional Neural Network
DER	Distributed Energy Resources
DG	Distributed Generation
DHI	Diffuse Horizontal Irradiance
DL	Deep Learning
DNI	Direct Normal Irradiance
DNN	Deep Neural Network
DPVS	Distributed Photovoltaic Systems
DR	Demand Response
DSO	Distribution System Operators
EV	Electric Vehicle
GHI	Global Horizontal Irradiance
GIS	Geographical Information Systems
HP	Heat Pump
IESO	Independent Electricity System Operator
IQR	Interquartile Range
LSTM	Long Short-Term Memory
ML	Machine Learning
NN	Neural Network
NREL	National Renewable Energy Laboratory
PV	Photovoltaic

R-CNN	Region-Based Convolutional Neural Network
SAM	System Advisor Model
SVC	Support Vector Classification
SVM	Support Vector Machines
SVR	Support Vector Regression

Indices

ω	scenarios
H	hours
s	RES type
Y	years

Parameters

α	Power output temperature coefficient for solar panels [pu/°C]
τ	Temperature of the photovoltaic cell [°C]
τ_{std}	Temperature of the solar cell under standard testing conditions [°C]
φ	Performance reduction factor of solar panels [pu]
$\hat{L}$	Estimated natural active power consumption [MW]
$\widehat{PV}$	Estimated PV power generation [MW]
G_{std}	Solar irradiance at standard conditions [kW/m ²]
$G_{Y,H}$	Global solar irradiance [kW/m ²]
$I_{Y,H}$	Total installed capacity of solar panels [kW]
m	Test data set size
N	Net load profile [MW]
n	Train data set size

Variables/Functions

σ	Sigmoid activation function [pu]
b	Biases [pu]
C	Confidence scores for test [pu]
c_t	Cell state [pu]
F, g	Activation functions [pu]
f_t	Forget gate [pu]
h_t	Hidden state [pu]
i_t	Input gate [pu]
o_t	Output gate [pu]
P_s	Power produced by photovoltaic panels [kW]
r_i	Confidence scores for training [pu]
s_j	Confidence scores for test [pu]
t	Time step [pu]
W	Random weights [pu]
x	Input state [pu]

y Output state [pu]

I. INTRODUCTION

THE integration of Distributed Photovoltaics (PV) Systems (DPVS) into power grids presents a unique set of challenges, especially due to their typically behind-the-meter installations, which makes them largely invisible to utilities and system operators [1]–[3]. Unregistered rooftop installations or decommissioned systems that may not be recorded can lead to significant gaps in utilities' records, as we noticed in our interactions with our utility partners. This can have a significant impact at the transmission feeder level since these feeders cover large areas. Thus, incomplete utility records may not account for installations predating mandatory reporting, decommissioned systems, off-grid systems, or attempts to avoid permit fees [4], [5]. Additionally, inaccuracies in utility data can arise from manual entry errors or discrepancies between reported and actual system configurations. Limited accessibility to distribution level network data by transmission system operators further complicates efforts to obtain comprehensive information, especially if some operators are reluctant to share data with third parties. In fact, it has been determined that significant portions of total PV rooftop residential capacity have not been recorded, as is the case of 30% PV rooftop installations in the UK and 52% in Germany [4], [5]. Accurate knowledge of DPVS capacity and profiles is important for effective grid management and energy forecasting, yet current data-gathering approaches often yield imprecise results due to unauthorized or undocumented installations [6]. The absence of monitoring of these systems can lead to significant uncertainty in net demand and complicate the operation of power grids [7], [8]. This lack of visibility impedes the ability of network operators to ensure proper power, operation and efficient management of local energy resources, highlighting the need for improved methodologies to gauge and integrate these systems more effectively [9], [10].

Traditional monitoring and management techniques are often insufficient for addressing the complexities introduced by widespread Distributed Energy Resources (DERs) [10]. This has spurred the development of load disaggregation techniques to better understand the performance of these systems and mitigate their impacts on the grid [11], enhancing the visibility of aggregated DERs behavior and facilitating more sustainable and resilient grid operations [8]. This paper addresses these challenges by proposing Machine Learning (ML) and Deep Learning (DL) methods for estimating the capacity and disaggregation of distributed PV systems, thus enhancing overall grid stability and facilitating energy management.

Numerous methods have been developed to estimate the generation capacity and real-time production of solar systems. One common approach involves the use of satellite and aerial imagery to detect solar PV systems and assess their physical characteristics such as size, tilt, and orientation [5], [12]. While this technique offers a broad estimate of PV installations' peak capacity, it falls short of accurately predicting real-time solar generation. DL models may misclassify objects such as solar thermal systems, skylights, or roads as PV systems, or fail

to identify PV systems due to factors such as low-resolution images and the large size of the areas being considered. Another common strategy for estimating PV output is model-based, utilizing localized irradiance measurements or estimations of beam, diffuse, and global solar irradiance derived from satellite data [13], [14]. These methodologies, although effective in certain contexts, often lack the precision needed for comprehensive real-time monitoring and analysis, particularly when dealing with diverse environmental conditions and large geographical areas. Hence, a combined approach integrating multiple data sources and advanced DL techniques including human annotation may enhance the accuracy and reliability of PV capacity identification.

A. PV Capacity Identification

The use of Geographical Information Systems (GIS) for the automated identification and capacity estimation of solar PV systems across extensive regions is gaining significant traction [5]. Advances in computer vision, specifically through semantic segmentation models and DL technologies, have enhanced the accuracy and efficiency of these identification processes [15]–[17]. Previous studies have demonstrated the successful use of GIS for detecting various structures, including buildings, vehicles, roads, and rooftops [5], [18], [19]. Although research specifically targeting the identification of solar PV systems is relatively limited, the existing results are promising [20]. Despite these advancements, challenges persist in identifying solar PV systems over large areas due to the need for high-resolution imagery (typically 0.3 meters per pixel or greater) and the difficulty in distinguishing solar panels from other similar structures, such as solar hot water systems, skylights, roads, and pools [5]. This underscores the need for continued refinement of these methods to enhance their accuracy and reliability.

Some studies have explored the capacity estimation of individual DPVS using advanced object detection algorithms. For instance, [21] proposes an object detection algorithm that utilizes high-resolution color satellite imagery to automatically identify the placement and dimensions of small-scale PV systems, with notable limitations such as the need for manual annotations of rooftop PV locations and rough capacity estimate results. In [22], the authors apply a U-Net network for identifying buildings in Wuhan, China, from satellite imagery, using empirical coefficients to estimate PV generation potential. Similarly, [23] utilizes Convolutional Neural Networks (CNNs) to identify PV panel installation sizes and locations in Switzerland from aerial photos, while [24] focuses on analyzing PV panel image characteristics to develop an ML model. Like [25], these papers focus on rooftop spaces on a city-wide scale utilizing computer vision techniques and rely on simple coefficients for calculating PV panel areas. Despite advancements, these methods generally do not address the variability of PV panel capacities, configurations, and locations, which requires more refined and comprehensive techniques, as discussed in this paper.

B. PV Power Generation Identification

Integrating a high penetration of DPVSs into the grid significantly impacts Demand Response (DR) capacity estimation (e.g., [26]), customer baseline load estimation (e.g., [27]), load forecasting (e.g., [28]), and distribution network planning (e.g., [29]). However, power output, especially those installed behind the meter, often remains invisible to utilities and system operators, making accurate information challenging to obtain due to various factors such as net metering. Additionally, while metering for larger PV installations is typically mandated, these meters often provide low-resolution data (e.g., 15-minute intervals) and are not required for smaller residential systems, further complicating the identification of DPVS contributions.

Several studies have addressed the challenges of load disaggregation for PV power identification. For example, in [8], various approaches such as online learning algorithms, Dynamic Fixed Share, and Linear Regression models to tackle these issues. Although these methodologies provide promising solutions, they remain constrained by the limitations of existing metering infrastructure. A notable identification approach involves a dual-phase estimation of DPVS power generation utilizing Support Vector Machine (SVM) based on customer load profile characteristics [1], underscoring its potential to improve DPVS profile estimation accuracy. These approaches highlight the need for continuous refinement to enhance the accuracy and reliability of load disaggregation in the context of DPVS identification.

Finally, studies on load disaggregation and DPVS profiling have focused on distribution systems [8], [30]. However, there is a lack of research at the transmission level, which is needed for power grid planning and operation, and is the focus of the current paper.

C. Contributions

A detailed literature review reveals the following knowledge gaps in existing studies on PV disaggregation:

- **Dependence on PV Historical Data:** Many PV disaggregation models rely on large amounts of historical data for distributed power sources and loads to train supervised learning algorithms. However, mostly aggregated data are available to transmission system operators, limiting the scalability and applicability of these models.
- **Transmission-level Focus:** Most existing studies concentrate on distribution-level PV disaggregation, with limited studies on transmission-level applications. However, as renewable energy integration increases, transmission grids play a crucial role in balancing supply and demand, ensuring inter-regional power exchanges, and maintaining grid stability. In this context, addressing distributed PV disaggregation at the transmission level is essential for utilities to enhance power grid planning and operation.
- **Scalability:** Current PV disaggregation methods often struggle to scale effectively across large systems, especially when dealing with diverse voltage levels. Given the increasing penetration of DERs, scalable solutions are needed to support real-time monitoring and decision-making.

- **Diverse Data Sources:** Limited research explores the integration of multiple data sources, such as satellite imagery and weather data, which could enhance the accuracy and robustness of PV disaggregation models. Integrating diverse data streams is important for improving model generalizability and ensuring reliable PV power estimation in large-scale power networks.

Based on the aforementioned discussions, this paper focuses on reporting the results of a collaborative project between the University of Waterloo and the Independent Electricity System Operator (IESO) of Ontario, Canada. The primary objective is to create a methodology for estimating PV power generation and net demand load at the transmission level, solely utilizing distribution grid feeder measurements at the transmission level to disaggregate DPVS generation and net demand. In the presented study, it is assumed that all visible PV systems identified through satellite imagery are operational. However, it is clear that the operating status of solar panels cannot be verified solely through satellite observations. Consequently, this assumption introduces a potential source of estimation uncertainty that may be addressed using additional data sources, such as smart meter readings, which were not available for the studies presented here and it cannot be assumed to be readily obtainable. Thus, the following are the main contributions of the paper:

- The use of satellite imagery and image-based deep learning techniques to estimate PV capacity, using GIS tools for precise panel detection and capacity estimation.
- A generalizable and scalable method for PV disaggregation from net load data, designed to function effectively in scenarios where other data sources are unavailable, unreliable, or not properly and timely updated.
- The application of synthetic data generation and multiple ML/DL models to predict PV power generation for net-load disaggregation.
- The use of actual transmission feeder-head power measurements to disaggregate PV generation and demand load downstream, considering multiple residential and grounded PV farms at the substation.

The rest of this paper is structured as follows: Section II outlines the background and framework that underpins the proposed techniques. Section III details the problem formulation, including the approaches for detecting PV panels based on GIS and for the estimation of solar PV generation. The load disaggregation techniques based on various ML/DL approaches used here are also described in this section. Section IV discusses the input data used, demonstrating the approach for PV panel detection and estimation of PV power generation in the Waterloo Region of Ontario, Canada, and presents the results of applying the developed load disaggregation methods in this region. Last, Section V highlights the main content and conclusions of the paper and discusses possible future research directions.

II. BACKGROUND REVIEW

A. Graphical Information Systems (GIS)

Satellite imagery plays an important role in GIS, offering a comprehensive means to acquire data that allows observing and analyzing relevant earth surface features [4]. Overlaying these images onto maps using GIS tools facilitates the quantitative analysis of certain regions of interest for the identification and classification of specific features, which is essential for detecting PV panels on rooftops and estimating their capacity [5], [18], [19]. The accurate identification of existing solar PV systems can be accomplished using advanced ML and DL image-based techniques [31], which enables precise and efficient renewable energy planning and management through a reliable and comprehensive estimation of solar energy distribution and capacity. Thus, these tools are used for the detection and capacity estimation of existing solar PV panels in this work, as explained in more detail in the next section.

B. Net-Load Disaggregation using ML/DL Techniques

Transmission system operators struggle to accurately estimate solar energy generated by PV systems from net-load measurements, particularly when there is limited information about PV generation [8]. Load disaggregation techniques allow the user to effectively address this issue, by decomposing combined load and PV generation data into distinct components. Thus, these methods can be used to estimate both the actual system demand and the contribution of PV panels (e.g., [32]). In systems with numerous unmonitored PV installations spread along a feeder, as illustrated in Fig. 1, disaggregating PV power is especially beneficial for system operators. Therefore, this study focuses on employing advanced ML-based algorithms and models to analyze aggregate power measurements at the transmission system level to estimate PV generation output from net-load measurements [8].

Modeling net-load disaggregation as an optimization problem allows the determination of the aggregate PV active power generation alongside a substation's active power load [8]. The primary goal is to find the optimal values that minimize the discrepancy between modeled and actual values while adhering to specific constraints. Incorporating ML and DL techniques significantly enhances the accuracy and reliability of these models. Hence, this is the approach used here to identify solar PV generation and load demand from actual net power metered values at High Voltage (HV)/Medium Voltage (MV) transformers in transmission systems, as explained in detail in the next section.

III. PROBLEM FORMULATION

A. General Approach

Feeder PV-load disaggregation at HV/MV substations involves separating the total solar PV generation in the feeder from the overall active power net-load measurements at the HV/MV substation. This methodology utilizes historical transmission-level data of net-load power measurements, along with accessible meteorological data to assess PV generation

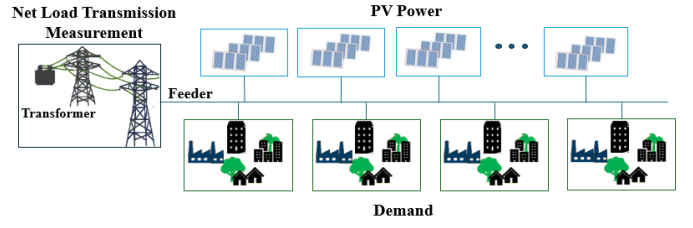


Fig. 1. PV generation identification at the transmission system level from net-load measurements in power grid feeders.

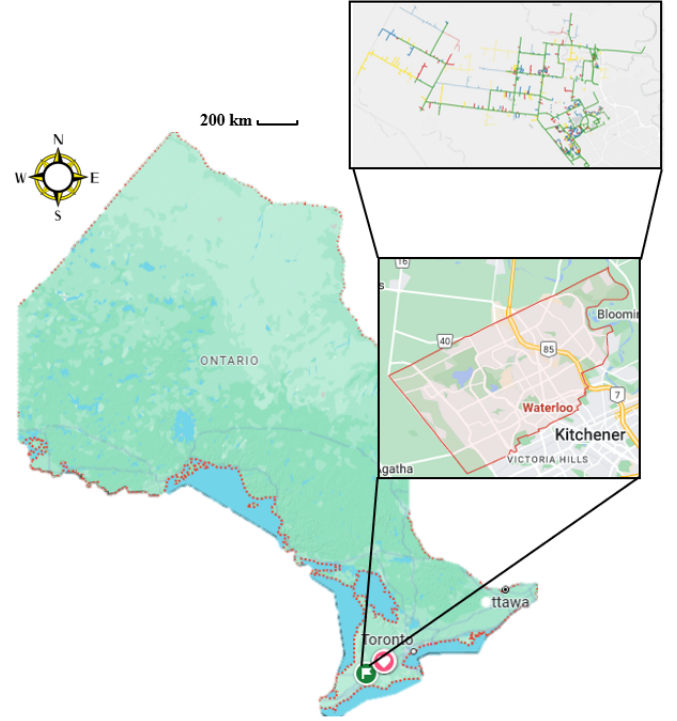


Fig. 2. Waterloo Region in Ontario, Canada.

from installed PV capacity estimates through the disaggregation of load profiles and solar power generation data.

At the transmission system level, there is typically no data available providing information on PV installed capacity below certain thresholds (e.g., 5 MW in Ontario, Canada [33]), with small PV panels dispersed across a wide service area associated with an HV/MV substation, featuring diverse configurations such as tilt angles and azimuths. To address this issue, estimates are needed of the PV capacity connected to the transmission system substation transformers of the region of interest, as illustrated, for example, in Fig. 2. This can be accomplished, as previously mentioned, using GIS and image DL-based techniques to detect and quantify the number of PV panels, allowing for an assessment of the overall capacity of micro-generation PV systems integrated into the distribution network serviced by HV/MV substations.

To estimate the PV generation based on the estimated installed PV capacity, synthetic PV power profiles can be generated using simulation methods. Despite variations in the tilt angles and orientation of the PV panels, these generally align with the region's latitude to maximize energy harvest

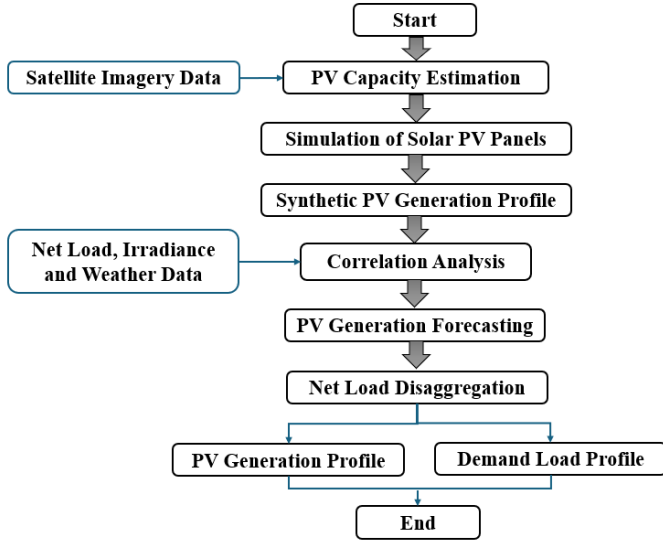


Fig. 3. Overview of the proposed PV generation and demand disaggregation approach.

[19]. Therefore, a virtual equivalent PV station can be assumed to estimate the aggregate output of all solar PV installations connected at the transmission system substation, based on the total estimated PV panel capacity and an assumed tilt angle and orientation, along with irradiation and weather data. This allows obtaining an aggregated PV output profile using an existing simulation tool. Based on this synthetic PV profile, supervised ML/DL models can be applied to disaggregate demand and PV profiles from net-load metered data.

The overall proposed procedure is depicted in Fig. 3, which is divided into the following four main phases, explained in detail next:

- 1) Estimation of the equivalent PV capacity and other relevant parameters based on GIS and simulation tools to produce synthetic PV generation aggregated profiles for forecasting purposes.
- 2) Gathering relevant data and performing a correlation analysis to determine pertinent features for PV generation forecasting.
- 3) Forecasting of aggregated PV generation based on synthetic PV output profiles and identified relevant features
- 4) Disaggregation of net-load power measurements to estimate PV power generation and load demand.

B. Solar PV Output Estimation

To estimate the PV generation capacity within the transmission system of a region, satellite imagery combined with sophisticated DL techniques can be used, as previously mentioned. Thus, image segmentation can be employed to precisely identify and delineate objects within an image. This process includes state-of-the-art object recognition techniques for locating and classifying objects with bounding boxes and semantic segmentation for pixel-level categorization [34], [35], which is particularly effective for detecting PV panels. Various segmentation models are available, each optimized for different types of imagery, study objectives, and data availability.

Hence, a mapping approach is used here that leverages pre-trained Deep Neural Networks (DNNs) for object detection, with human annotations for validation to ensure accuracy.

For object detection, the DNN-based Mask R-CNN algorithm in [34], [35] is used here as the pre-trained object detection model. As illustrated in Fig. 4, Mask R-CNN is a specialized DNN designed to process grid-like data, such as images represented as a two-dimensional pixel array. It processes the data in two phases to predict object classes and their bounding boxes. In the initial phase, two NNs, namely, the backbone and the Region Proposal Network, suggest potential object bounding boxes. In the subsequent phase, features are extracted from each proposed box, followed by classification and segmentation mask prediction for each region of interest. Hence, based on satellite images for the region of interest, a pre-trained Mask R-CNN pre-trained panel segmentation model can be used to identify PV panels in the region. The raw prediction output provides a list of bounding boxes with the following confidence scores, C_{train}^i and C_{test}^j [36]:

$$\begin{aligned} \{C_{\text{train}}^i &= (r_1^i, r_2^i, r_3^i, \dots) \mid i = 1, 2, \dots, n\} \\ \{C_{\text{test}}^j &= (s_1^j, s_2^j, s_3^j, \dots) \mid j = 1, 2, \dots, m\} \end{aligned} \quad (1)$$

where n and m represent the number of training and testing image tasks, respectively, and r_k^i and s_k^j correspond to the confidence scores between 0 to 1 for training and testing, respectively, of prediction object k . To assess the accuracy of the results, manual annotations of all solar PV arrays in the aerial images were conducted to validate and enhance the results of the algorithmic detections with the help of human annotations.

The identification of solar PV systems is influenced by a variety of factors that pose significant challenges, as expected with DL-based models when applied to complex, real-world pictures on a large scale. Thus, the following key difficulties that need to be considered in this process are:

- False Positives: Solar PV systems are susceptible to misclassification due to their resemblance to various other objects. In the current study, such misclassifications include shadows from nearby structures, dark-colored roofing materials, reflective surfaces on roofs, water pools, skylights, vehicle windshields, building edges, and car roofs. Some representative examples of these instances are illustrated in Fig. 5. This issue can be effectively addressed with human annotations.
- False Negatives: Identifying solar PV systems can be challenging, even for experienced annotators, particularly when black panels are installed on dark rooftops or when image resolution is insufficient. In instances where manual annotation is difficult, as in the case of small or irregularly shaped PV installations, model predictions tend also to be prone to errors, resulting in reduced classification or segmentation performance, making detection more complex. Environmental factors such as snow cover and panel soiling can also obscure visibility, further increasing the likelihood of missed detections [5]. This issue can be minimized by being more careful during

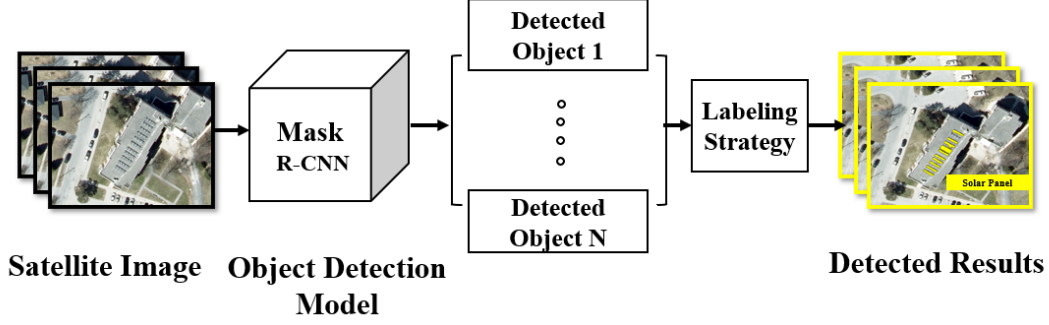


Fig. 4. Mask R-CNN architecture for detecting and segmenting PV solar panels.

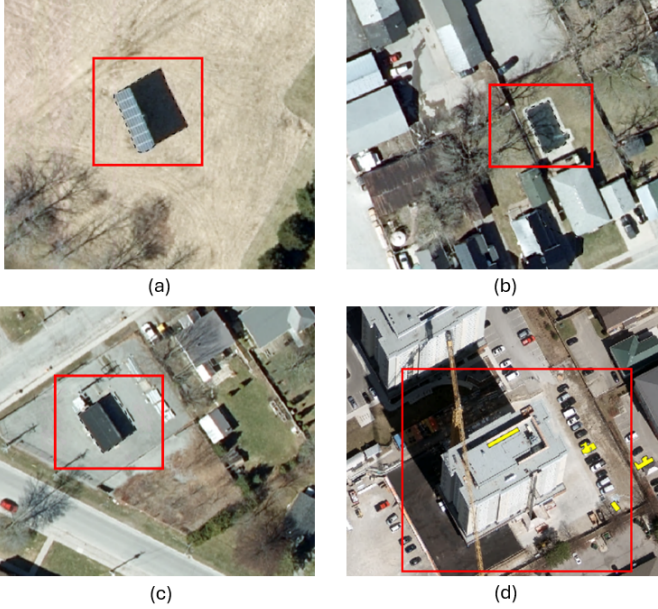


Fig. 5. Example of false positive segmentation of solar PV panels: (a) shadows from surrounding structures, (b) water pool, (c) dark-colored roofing materials, (d) skylights and car roofs.

the human annotation process, but it cannot be fully eliminated and is hence a potential source of error.

- **Consistency of Imagery Datasets:** Overhead imagery is captured from specific angles and times of day. As a result, images within the same dataset may exhibit differences in shadow presence and size, visibility of vertical walls, or other features that vary with the camera angle. These inconsistencies can be addressed by ensuring that a sufficiently large set of images is annotated. However, in some cases, the significance of the differences across images is unclear within the dataset [5], making these errors difficult to detect.

The power output of the detected PV panels $P_{sY,H}$ can be calculated as a function of hourly incident solar irradiance, PV cell temperature, and a derating factor that considers issues such as dust accumulation, wire losses, and other deviations from optimal performance, under standard test conditions (stc), as follows [19]:

$$P_{sY,H} = \varphi I_{Y,H} \left(\frac{G_{Y,H}}{G_{stc}} \right) [1 + \alpha(\tau_{Y,H} - \tau_{stc})] \quad \forall Y, H \quad (2)$$

where φ is the derating factor of solar panels; $I_{Y,H}$ is the total installed capacity of solar panels; $G_{Y,H}$ is the incident solar irradiance at a specific time and location; G_{stc} is the standard test conditions irradiance; α is the temperature coefficient of power for solar panels; $\tau_{Y,H}$ is the actual solar cell temperature; τ_{stc} is the solar cell temperature at standard test conditions; Y is the year; and H is the hour of the prediction. This formula is integrated into the practical and reliable System Advisor Model (SAM) open-access tool for generating PV output data developed by the National Renewable Energy Laboratory (NREL) [37], which uses real solar irradiance and weather data to accurately simulate PV output. The synthetic dataset generated by SAM not only includes the PV output $P_{sY,H}$, but also Direct Normal Irradiance (DNI), Diffuse Horizontal Irradiance (DHI), ambient air temperature, calculated Global Horizontal Irradiance (GHI), wind speed, and dew point, which can then be used for PV forecasting purposes.

C. Solar and Load Power Disaggregation

The goal of the solar and load disaggregation method is to distinguish between net load and solar power. This can be accomplished using a supervised learning model that requires net-load measurements and PV generation data during the training phase. These data are considered to be collected from the HV/MV substation of interest, which includes numerous unmonitored residential PV systems. Based on the estimated equivalent capacity and power of the installed PV system, a comprehensive method is proposed next to disaggregate solar generation and load demand.

To accurately separate $\widehat{PV}$ forecasted generation from the estimated demand $\widehat{N}$, the net-load profile $L = \widehat{N} - \widehat{PV}$ can be used.

For this purpose, the approach depicted in Fig. 6 is applied using ML/DL techniques for PV profile forecasting. This disaggregation method for net-load component disaggregation leverages net-load data from the HV/MV substation, synthetic PV data, and meteorological information used for forecasting solar PV generation. The forecasting can be accomplished through the application of ML/DL time series algorithms,

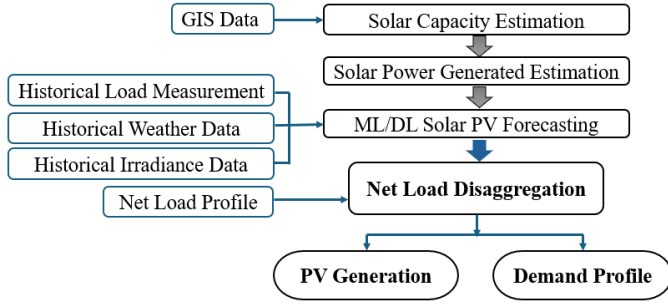


Fig. 6. Proposed approach for disaggregating net-load components.

including Long Short-Term Memory (LSTM), CNN-LSTM, Bidirectional LSTM (BiLSTM), SVM, K-Nearest Neighbors (KNN), and Regression methods.

D. Machine Learning Models

Linear Regression, KNN, and SVM are benchmark ML techniques that have been employed for both time series forecasting and disaggregation tasks [8], [17], [32], [34].

1) Regression Model

Linear Regression is a simple yet powerful statistical method that models the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data. In the context of time series forecasting, Linear Regression can be employed to predict future values based on past trends, offering a straightforward approach to understanding and predicting time-dependent data. It is computationally efficient, but it cannot capture nonlinear patterns [38].

2) KNN Model

KNN is a non-parametric, instance-based learning algorithm that predicts the output based on the majority vote of its nearest neighbors in the feature space. It is particularly useful for classification and regression tasks, including time series forecasting, by utilizing historical data points to predict future values. This algorithm bypasses complex equation solving by leveraging raw data and the correlation between the target point and its nearest neighbors. In the current study, KNN is utilized to identify historical time series features that resemble future PV generation, making it particularly effective for preprocessing time series data [39].

The KNN training process is based on measuring how similar or different two data points are in a multi-dimensional space in terms of an Euclidean distance norm. The nearest neighbors are then selected based on their similarity, ensuring that predictions are grounded in closely related past observations [40]. The initial values of the KNN model hyperparameters can be chosen based on typical values found in the existing literature.

3) SVM Model

SVM are supervised learning models used for both classification and regression. In time series applications, SVM works by finding the optimal hyperplane that maximizes the margin

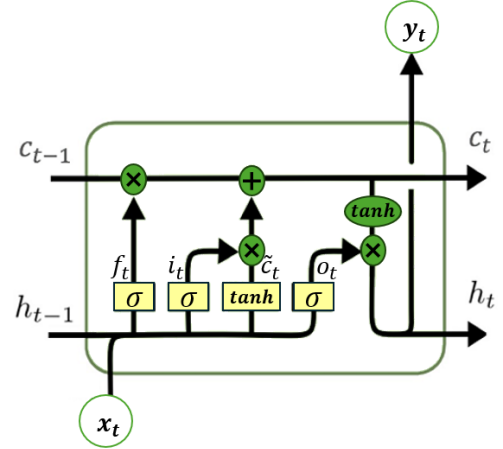


Fig. 7. Structure of LSTM cell with input x_t and output y_t [34].

between different data points, ensuring robust predictions by minimizing error. SVM is particularly effective in handling non-linear data through the use of Kernel functions.

For PV power forecasting, SVM regression is utilized here to model the relationship between input features, such as weather parameters (temperature, irradiance, wind speed, etc.), and the output, i.e., generated PV power. The goal in SVM regression is to find a predictive function that closely matches actual power output [41]:

When dealing with non-linear dependencies, SVM employs Kernel functions to transform the input data into a higher-dimensional space where patterns become easier to separate. Popular Kernels include polynomial, Gaussian radial basis function (RBF), and hyperbolic tangent [42]. This ability to map complex relationships makes SVM particularly effective for forecasting PV power, which is influenced by fluctuating and interdependent environmental factors.

E. Deep Learning Models

1) LSTM Model

LSTM is a type of Recurrent Neural Network (RNN) designed to overcome the vanishing and exploding gradient issues that standard RNNs experience during training, thereby enhancing their ability to handle long-term dependencies in sequential data. LSTMs utilize a unique architecture featuring memory cells and gates to store, update, and retrieve information over extended time intervals. The LSTM's ability to store and retrieve long-term memory is crucial for modeling complex time-dependent patterns, such as those found in PV power generation [34].

As illustrated in Fig. 7, an LSTM block consists of several essential components: x represents the input state, y denotes the output state, t is the present time step, and σ stands for the sigmoid activation function. The core elements include the input gate i_t , cell state $\tilde{c}_t$, forget gate f_t , hidden state h_t , and output gate o_t . The structure and functionality of an LSTM cell can be defined as follows for random weights W_f , W_i , W_c , and W_o and biases b_f , b_i , b_c , and b_o [34]:

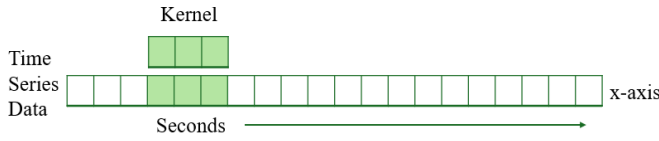


Fig. 8. 1D-CNN Computational process.

$$f_t = \sigma(W_f[h_{t-1}, x_t]^T + b_f) \quad (3)$$

$$i_t = \sigma(W_i[h_{t-1}, x_t]^T + b_i) \quad (4)$$

$$\tilde{c}_t = \tanh(W_c[h_{t-1}, x_t]^T + b_c) \quad (5)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \quad (6)$$

$$o_t = \sigma(W_o[h_{t-1}, x_t]^T + b_o) \quad (7)$$

$$h_t = o_t \odot \tanh(c_t) \quad (8)$$

The σ and $\tanh$ activation functions are used to enable selective memory updates and transformations. Additionally, key hyperparameters such as the number of neurons, learning rate, and dropout rate are fine-tuned to optimize the model's performance and avoid underfitting or overfitting [34].

2) CNN-LSTM Model

CNN-LSTM is a new hybrid approach that combines the strengths of CNNs and LSTM networks to enhance the accuracy of time series forecasting [32], [34]. CNN layers are employed to automatically extract local patterns and non-linear relationships from the input data, which is particularly beneficial for capturing complex, multi-dimensional features in photovoltaic PV generation patterns. The extracted features are then passed to the LSTM layers, which are well-suited for modeling temporal dependencies and learning long-term relationships in the data. One-Dimensional CNNs (1D-CNNs) are particularly useful for processing sequential data such as time series or natural language, where the Kernel (or filter) moves along the input data in one dimension, as shown in Fig. 8. This allows for the capture of local dependencies and extract features across the sequence, making it well-suited for applications where data points are temporally or spatially ordered, such as PV power generation time series [43].

The CNN model is based on a DL architecture commonly used for processing sequential data, which extracts important features from the input data by applying convolutional layers followed by pooling layers. In the context of PV power generation, a CNN can be used to classify weather conditions based on input patterns. The CNN architecture typically involves convolution layers with ReLU activation functions and fully connected layers to map extracted features to the output labels [43].

The CNN-LSTM model applied here uses CNN for weather classification, and based on the classified weather condition, an LSTM model is selected to forecast PV power generation. This hybrid approach allows the model to handle different weather conditions separately, reducing the complexity and variability of data fitting. The overall structure of the CNN-LSTM hybrid model facilitates capturing non-linear relationships and temporal dependencies, offering better forecasting accuracy than a standalone LSTM model. The use of a CNN

for feature extraction and LSTM for sequential forecasting, as shown in Fig. 9, allows the hybrid model to significantly reduce forecasting error, providing a more reliable prediction of PV power generation.

Compared to a standard LSTM model, the CNN-LSTM model demonstrates a superior ability to capture local trends and non-linear relationships between diverse features and PV generation patterns. This results in fewer forecast errors and a closer match between actual and predicted values. By leveraging the combined advantages of CNN and LSTM, the CNN-LSTM model provides more accurate forecasts, making it a valuable tool for improving prediction and energy management.

3) BiLSTM Model

The BiLSTM model in this study is designed to efficiently process time series data by leveraging bidirectional temporal contexts. A sliding window technique can be employed during training to segment the input sequence, enabling the model to extract meaningful temporal features.

Fig. 10 shows the structure of BiLSTM. While standard LSTM captures information from previous time steps through a backward-propagation path, BiLSTM enhances this capability by adding a forward-propagation path. This forward LSTM layer processes data in the reverse direction, enabling the model to consider both past and future information simultaneously. As a result, the output layer of BiLSTM integrates these bidirectional contexts, enabling the hidden cells to utilize comprehensive temporal information during model training. Modern BiLSTM models incorporate a sliding window technique applied to the input sequence [44]. The LSTM block structure forms the core of the BiLSTM model proposed here.

Modern BiLSTM models incorporate a sliding window technique applied to the input sequence [44]. The LSTM block structure forms the core of the BiLSTM model proposed here. In these models, the computational cost of the forward and backward layers is equivalent. However, the direction of hidden state propagation is reversed in the forward layer, allowing the model to capture both past and future time-dependent information. The mathematical formulation for the forward and backward LSTM layers of the BiLSTM network can be expressed as follows [45]:

$$h_f = F(W_{f1} x_f + W_{f2} h_{t-1}) \quad (9)$$

$$h_b = F(W_{b1} x_t + W_{b2} h_{t+1}) \quad (10)$$

where h_f and h_b are the outputs of the forward and backward LSTM layers, respectively, as depicted in Fig. 10, and F is an activation function such as σ or $\tanh$. The final output of the BiLSTM model is then computed by combining these outputs as follows [45]:

$$y_i = g(W_{o1} h_f + W_{o2} h_b) \quad (11)$$

where g is another activation function.

To improve performance, the input data can be normalized, and the sliding window length can be tailored to the specific characteristics of the time series data. The optimal window size can then be determined by analyzing the data's periodicity,

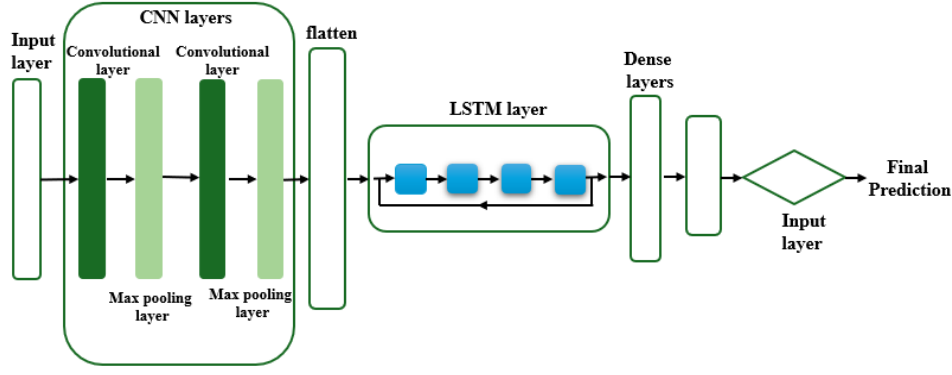
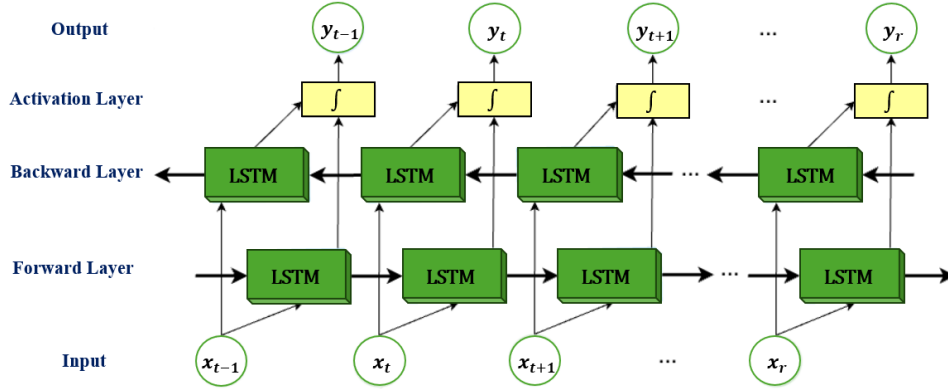


Fig. 9. CNN-LSTM hybrid model.

Fig. 10. Structure of BiLSTM cell with time series inputs x_t and outputs y_t [44].

temporal dependencies, and the inherent patterns within the time series. By adjusting the window length to suit the underlying structure of the data, the model can extract more meaningful temporal features and improve predictive accuracy [44].

Correlation analysis is needed to identify historical weather data and other relevant feature values that are critical for accurate prediction. Based on this analysis, the most significant features can be selected as inputs for the various ML/DL forecasting techniques evaluating their respective error. By obtaining the PV power generation profile through these models, the aggregated demand load profile can be calculated. The disaggregated profiles, i.e., N and $\widehat{PV}$ profiles, can then be utilized for proper system operation.

IV. RESULTS AND DISCUSSION

A. Data Sets

1) Transmission Level Measurement

The first dataset used consisted of the transmission-level net-load data supplied by the IESO, which corresponded to two HV/MV transformers in the Waterloo Region illustrated in Fig. 2, with a rate of 108 MW. These data span the period from January 2018 to December 2021 and include measurements recorded with 1 min granularity, which were aggregated to the 15 min interval used here.



Fig. 11. Solar PV panels segmentation in the Waterloo Region.

2) GIS Data

The second dataset for the PV panel detection task are images that span a total area of 240 km² in the Waterloo Region. The images used here, which are not publicly available, were provided by the Geospatial Center at the University of Waterloo. All imagery corresponded to aerial photography in April 2021 including images from the Landsat 8 satellite. Large satellite images, ranging from 1 km² pixels, were cropped to create patches suitable for semantic image segmentation as shown, for example, in Fig. 11, with solar PV panels annotated in yellow. The implementation of the pre-trained Mask R-

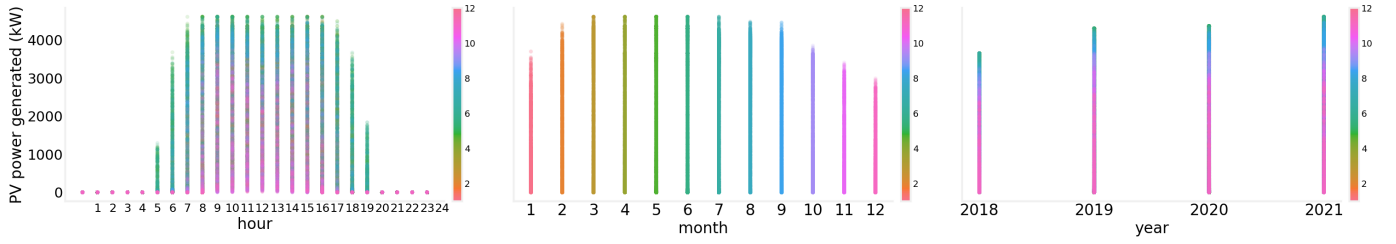


Fig. 12. Simulated PV power output from SAM software with the colors signifying the probability of the PV output values, with red for more likely values and green-blue for least likely.

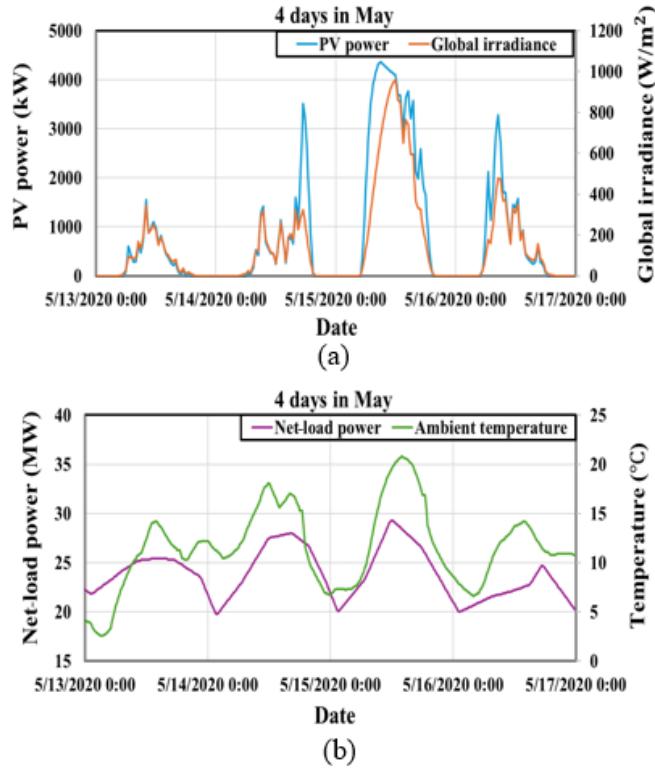


Fig. 13. Time series for four days showing PV power generation and net-load versus solar irradiance and ambient temperature; (a) and (b) for May 2020.

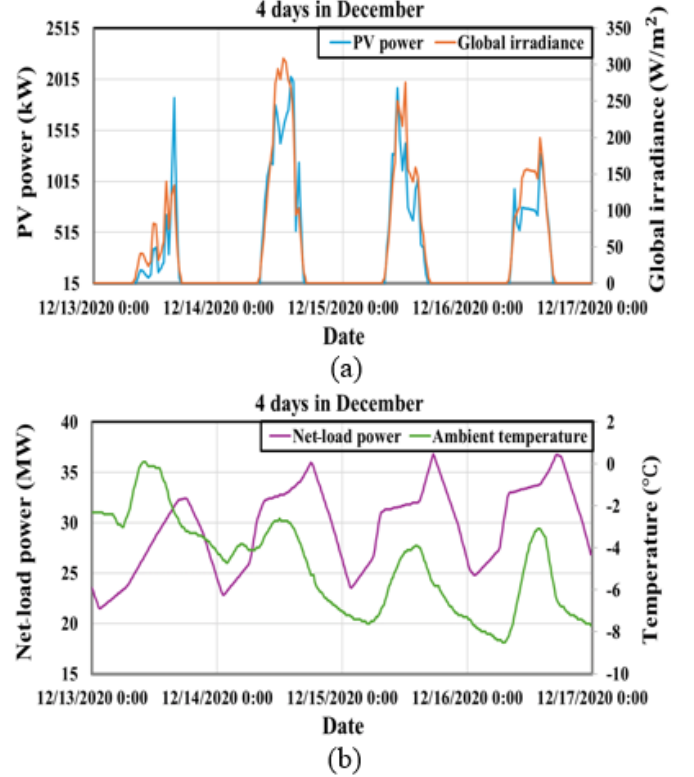


Fig. 14. Time series for four days showing PV power generation and net-load versus solar irradiance and ambient temperature; (a) and (b) for December 2020.

CNN approach was carried out using Python's ArcGIS Pro API applied to the available high-resolution imagery of the region of interest, and a pre-trained model using a large dataset of color aerial imagery from Fresno, CA, USA [31].

3) PV Dataset

A synthetic PV data generation method were implemented using the SAM software [46]. Fig. 12 illustrates the PV power output values and the likelihood obtained for hours and months across the different years of study.

4) Meteorological Data

The 30-minute meteorological datasets used were obtained from the NREL [37], for the Waterloo Region. The dataset includes variables such as temperature, irradiance, wind speed, dew point, and others, which were used to identify relevant correlations for solar generation forecasting revealing how variations in meteorological conditions influence solar energy output.

5) Temporal Features

These consist of 15 min intervals for the 24 h a day, the months of the year, the day of the week, and years. The synthetic PV data depicted in Fig. 12 revealed that both the hour and month significantly affect the PV output.

6) Data Preprocessing

The data obtained from various repositories had to undergo preprocessing before being used for forecasting and disaggregation. This preparation involved several steps, including data cleaning, synchronization, and vectorizing (one-hot encoding) to produce a clean and well-structured dataset. The raw data contained high-frequency noise, missing entries, incorrect labels, and duplicate records. Missing values were addressed by appropriate imputation methods, such as averaging.

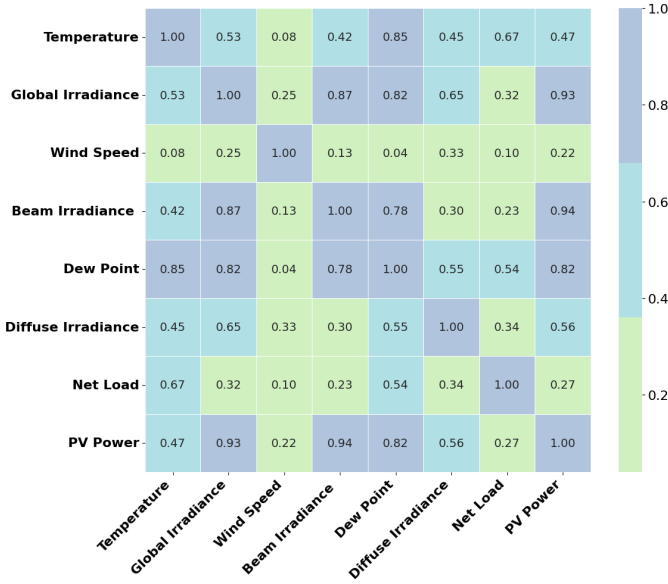


Fig. 15. Correlation metrics between solar power output and various factors in May.

B. Feature Extraction and Correlation Analysis

It is important to synchronize the data since they were collected at different sampling rates: 1 minute for load data, 30 minutes for weather data, and 15 minutes for irradiance data. In this study, all datasets were synchronized using their respective timestamps and resampled to a uniform 15-minute interval.

As previously mentioned, the proposed model utilizing supervised learning requires data for both net load (input vector) and estimated PV generation (label) from the feeder for the training process. The dataset is partitioned into 75% for training, 10% for validation, and 15% for testing. Training dataset was used to fit the model parameters, while the validation set provided feedback for model performance during hyperparameter tuning. The test set, reserved for final evaluation, ensured an unbiased assessment of the models' generalizability. Moreover, categorical variables were transformed into numerical values through one-hot encoding [47], [48], which created a new binary feature for each category, allowing categorical data to be represented in a format suitable for numerical analysis and model training.

Dimensionality reduction, also known as feature projection, was utilized to transform data into a new subspace, retaining essential information and reducing noise and redundant variables, to accurately represent the input signal and improve load identification by isolating valuable information in a smaller or similarly sized subspace [9]. Fig. 13 and Fig. 14 depict the relation between PV power generation and net load versus solar irradiance and ambient temperature for four days in May and December 2020, illustrating the nonlinear relation between solar irradiance and PV as per (2), derived from the PV synthetic data.

Fig. 15 illustrates the Pearson correlation coefficients between solar energy generation and various features for forecasting. A positive correlation is obtained when the coefficient

exceeds 0.3, showing that temperature, GHI, DNI, DHI, and dew point have significant impact on PV power generation. Furthermore, temporal factors such as time of day, day of the week, and month of the year are also important contributors. These pre-processed features were provided to the DL and benchmark ML models to ensure consistency across all models and standardize the input data. This guaranteed a fair and meaningful performance comparison between the two model types.

C. Mask R-CNN PV Detection

In semantic segmentation, the primary objective is to assess the accuracy of predicted and labeled masks. However, in object detection, the Intersection over *Union* (*IoU*) metric is often more meaningful than accuracy, especially in cases where small objects are present [49]. This metric evaluates the overlap between predicted and ground truth masks. Since PV panels occupy only a small fraction of the image, accuracy may remain high even if many panels are missed. In such cases, the following *IoU* metric provides a more reliable evaluation of model performance [50]:

$$IoU = \frac{AreaofOverlap}{AreaofUnion} = \frac{|A \cap B|}{|A \cup B|} \quad (12)$$

The proposed methods attained a performance range of 65%-70%, as determined by this metric.

In some instances, the *IoU* metric does not provide insights into the completeness of the model, in which case this can be complemented with additional metrics, including *Accuracy*, *Precision*, *Recall*, and *F1Score*. Thus, *Accuracy* can also be used to assess the DL model's performance with respect to human annotation, which is defined as follows [51]:

$$Accuracy = \frac{TP}{TP + TN + FP + FN} \quad (13)$$

where *TP*, *TN*, *FP*, and *FN* represent, respectively, the total number of true positives (correctly detected PV panels), true negatives (non-PV objects correctly classified as non-PV), false positives (non-PV objects incorrectly classified as PV panels), and false negatives (actual PV panels not detected by the model). In the PV detection task, these parameters are measured based on PV panel area in square meters. The model's reliability can be also assessed using the *Precision* metrics, which is computed as the ratio of correctly predicted PV panels (*TP*) to the total number of predicted positives as follows [51]:

$$Precision = \frac{TP}{TP + FP} \quad (14)$$

This is used to determine how often a positive prediction is correct. Finally *Recall*, which measures the completeness of the model, can be used to assess the proportion of labeled PV panel pixels that are correctly identified as follows [51]:

$$Recall = \frac{TP}{TP + FN} \quad (15)$$

These metrics allow a comprehensive evaluation of the model performance and are widely utilized in semantic segmentation

TABLE I
PERFORMANCE METRICS FOR MASK-RCNN MODEL EVALUATION

Metric	Percent
<i>IoU</i>	65%
<i>Accuracy</i>	99%
<i>Precision</i>	80.6%
<i>Recall</i>	65.9%
<i>F1score</i>	72.5%

tasks. However, to balance both *Precision* and *Recall*, the following *F1Score* can also be used as the Harmonic Mean of these two metrics [52]:

$$F1score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (16)$$

Table I summarizes the performance results of the proposed model for an area of 240 km² in the Waterloo Region.

D. PV Capacity Estimation

The estimation of PV power generation was conducted using GIS tools and DL image-based techniques applied to high-resolution satellite imagery to identify PV panels. Manual annotations of PV panels were conducted thrice to ensure consistency. In rare instances, the algorithm detected features that were initially overlooked during manual annotation. This back-and-forth process allowed the reliable identification of the PV panels in the region of interest. To validate the capacity estimates, data from Enova Power Corp. provided information on the maximum average capacity of solar panel installations in the Waterloo Region, including both groundmounted and rooftop PV systems from 2018 to 2021. This information is summarized in Table II, illustrating how the total solar capacity has increased significantly over the years, with 2021 representing a 32.69% increase compared to 2018. Thus, in 2021, the total solar panel capacity was recorded at 4570 kW, with the GIS-based estimated capacity closely matching it at approximately 4560 kW. Due to the unavailability of image data for the years 2018, 2019, and 2020, Enova's data were used as the capacity forecast. This assumption was made with the expectation that the GIS-based estimates would produce reliable capacity forecasts.

TABLE II
ACTUAL INSTALLED PV CAPACITY AT RELEVANT HV/MV
TRANSFORMER

Year	2018	2019	2020	2021
Total Capacity (kW)	3444	4240	4270	4570

E. Hyperparameters Settings

To achieve optimal performance in the proposed model, it is critical to carefully tune hyperparameters. The hyperparameters of the BiLSTM model, such as the number of neurons, activation functions, and learning rates, were fine-tuned based on a review of similar articles and iterative experimentation in the relevant literature. Thus, for the LSTM-based model, 128 nodes were selected to address potential issues such as

overfitting and underfitting [53]. Furthermore, early stopping was implemented to halt training when the validation loss did not plateau, effectively preventing overfitting and reducing unnecessary computational overhead. Selecting the appropriate number of nodes is important, as a higher number can lead to overfitting, with the model memorizing the training data but failing to generalize for unseen data, while a lower number may result in underfitting, with the model lacking the needed complexity to capture key patterns in the data. In this model, each node serves as a computational unit that processes inputs through a transfer function and connects them to outputs via weighted connections.

The ADAM optimizer was used for training all LSTM-based models, as it is a well-established optimization algorithm for deep neural networks. It leverages a weighted average to monitor the relative prediction error of the loss function and is based on the Stochastic Gradient Descents (SGDs) process [54]. A key advantage of ADAM is its ability to mitigate the effects of outliers while requiring only first-order gradients, making it computationally efficient. One notable feature of ADAM is its calculation of both momentum and the second moment of the gradient using exponentially weighted moving averages, which contributes to its robust convergence [55], [56].

An important hyperparameter in this study is the dense layer integrated into each model. This configuration was chosen to balance computational efficiency and model performance, ensuring an optimal runtime without compromising accuracy. Additionally, a learning rate of 0.001 was selected for all LSTM-based models, as this value has been shown to yield optimal results across various studies. The dense layer was strategically configured to ensure consistency and adaptability across different models.

Table III presents the key hyperparameters used in the LSTM, CNN-LSTM, and BiLSTM models. While certain hyperparameter values, such as the learning rate, are shared across all models, other parameters are unique to the specific architecture. For example, the ReLU activation function is used in the CNN-LSTM and BiLSTM models, whereas the LSTM model employs the tanh activation function. Additionally, the CNN-LSTM incorporated a Kernel size of 1, while the BiLSTM utilized two Kernel sizes, 1 and 2, to capture hierarchical temporal patterns effectively.

The dropout rate hyperparameter employed in LSTM was carefully chosen to improve model generalization, since this is commonly used to mitigate overfitting by randomly setting a portion of the network's weights to zero during training, ensuring better generalization for unseen data. The Padding strategy was also adapted to the specific model structure, using "Valid" Padding for the CNN-LSTM model and "Causal" Padding for the BiLSTM model. These adjustments were made to enhance each model's predictive accuracy and robustness. Additionally, the inclusion of Max-Pooling in the CNN-LSTM model further strengthened its ability to extract spatial features from input data, a feature not utilized by the LSTM and BiLSTM models.

As seen in Table IV, in addition to the LSTM-based models, several benchmark ML techniques were also considered, each

TABLE III
HYPERPARAMETER SETTINGS FOR LSTM, CNN-LSTM, AND BiLSTM MODELS

Hyperparameter	LSTM	CNN-LSTM	BiLSTM
Epoch	20, 30, 40	20, 30, 40	20, 30, 40
Batch size	100	100	100
Number of units	128	128	128
Dropout rate	0.2	N/A	N/A
Kernel size	N/A	1	1st=1, 2nd=2
MaxPooling size	N/A	1	N/A
Padding	N/A	Valid	Causal
Activation function	tanh	ReLU	ReLU
Dense layer units	1	1	1
Train data	75% dataset	75% dataset	75% dataset
Validation data	10% dataset	10% dataset	10% dataset
Test data	15% dataset	15% dataset	15% dataset
Optimizer	Adam	Adam	Adam
Learning rate	0.001	0.001	0.001

TABLE IV
HYPERPARAMETER SETTINGS FOR LINEAR REGRESSION, KNN, AND SVM MODELS

Techniques	Hyperparameter	Value
Linear Regression	None	N/A
	Lag count	3
KNN	N_neighbors	140
	Leaf size	Default 30
	Power parameter p	Default 2
	Lag count	3
SVM	Kernel	Radial Basis Function (RBF)
	C	1
	Degree	3
	Lag count	3

differing significantly in structure and implementation. Unlike LSTM models, which rely on sequential data processing and recurrent layers, these models leverage simpler mathematical and statistical approaches for analysis and prediction. However, due to their lack of consideration for temporal dependencies in the data, these models may be not as effective in comparison to the LSTM algorithm, which accounts for the complexity of sequential patterns [57].

To ensure a fair comparison, lag features for the SVM, KNN, and Linear Regression models were included. The hyperparameters for these models were carefully selected to optimize their performance according to the characteristics of the dataset. For Linear Regression, no additional hyperparameters were required, as this technique is based on a closed-form solution that minimizes the mean squared error between predicted and actual values, thus being a baseline model for regression tasks. The KNN algorithm was configured with 140 neighbors, which is based on similar values found in the existing literature plus a trial-and-error process, specifying the number of nearest data points considered when making predictions. The default leaf size of 30 was used, as it strikes a balance between computational complexity and performance of the search tree structure. For SVM, the Radial Basis Function (RBF) Kernel was employed, as it is well-suited for capturing non-linear relationships in the data, with the Kernel hyperparameter defining the transformation of input data into a higher-dimensional space, enabling linear separability. The C parameter, which regulates the trade-off between maximizing the margin and minimizing classification error, was set to 1. The Degree was fixed at 3, which is a value specific to polynomial Kernels, but it does not directly affect the RBF

Kernel. The Gamma parameter was set to “Scale” to adjust the influence of each individual training example, allowing the model to better capture the distribution of the data.

The aforementioned benchmark models served as a base to evaluate the LSTM-based DL models, providing a baseline and alternative perspectives of the data. Although the hyperparameters for these models are simpler than those of the LSTM-based models, their proper configuration remains important for achieving optimal performance.

F. Forecasting Error Metrics

As outlined in Section III, this study employs six ML/DL techniques, i.e., LSTM, CNN-LSTM, BiLSTM, SVM, KNN, and Regression, for forecasting and disaggregation. The evaluation criteria for these techniques are based on the following five key performance indicators [58]:

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i - x_i)^2}{n}} \quad (17)$$

$$MAE = \frac{\sum_{i=1}^n |y_i - x_i|}{n} \quad (18)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - x_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (19)$$

$$SD = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}} \quad (20)$$

$$PA = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \quad (21)$$

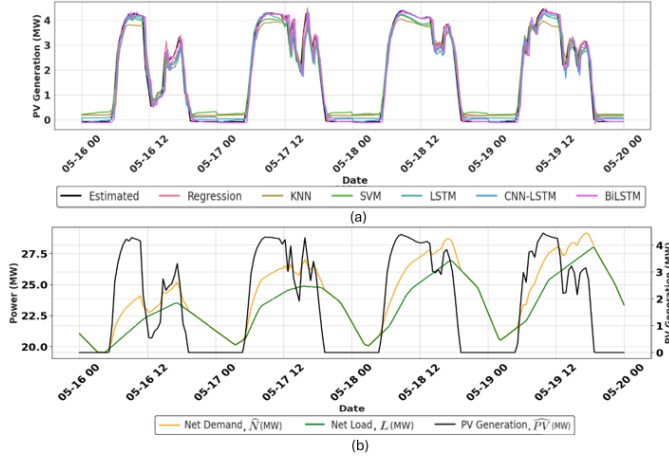


Fig. 16. Measured and disaggregated powers for May 2021: (a) PV generation forecast and (b) BiLSTM disaggregation results.

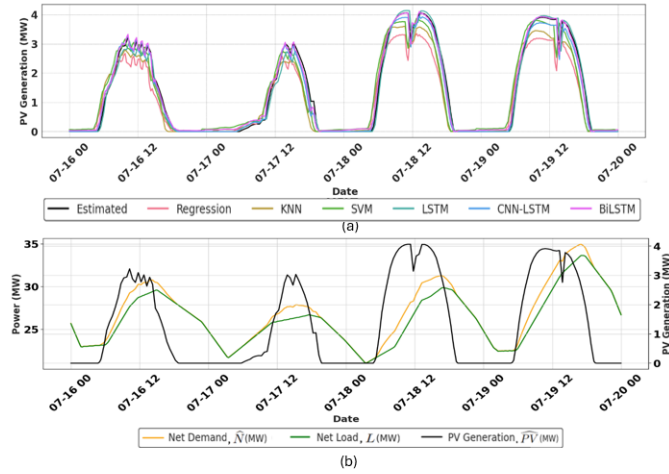


Fig. 17. Measured and disaggregated powers for July 2021: (a) PV generation forecast and (b) BiLSTM disaggregation results.

where y_i denotes the estimated PV value, $\bar{y} = \sum_{i=1}^n y_i/n$ represents the average mean of the estimated PV values, x_i is the predicted output, $\bar{x} = \sum_{i=1}^n x_i/n$ corresponds to the average mean of the predicted output, and n is the number of data points considered [44].

G. Application and Comparison

The estimation outcomes for PV generation and load demand disaggregation are depicted for May 2021 in Fig. 16 with similar results for July and December 2021 shown in Figs. 17 and 18, respectively. PV synthetic data from 2018, 2019, and 2020 were used for designing the ML/DL models, while estimated PV output power for 2021 was utilized as a reference for comparison purposes. The results for the Regression, KNN, SVM, LSTM, CNN-LSTM, and BiLSTM models are depicted in these figures, which show that the proposed BiLSTM method produced the best results in general, as discussed next based on the calculated performance indicators.

Table V compares performance indicators for 2021. To ensure a fair evaluation, the error metrics excluded time samples where PV generation is zero, thus preventing overestimation

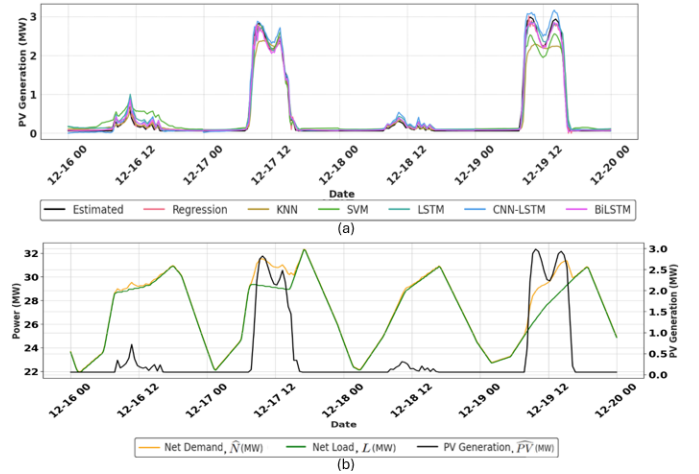


Fig. 18. Measured and disaggregated powers for December 2021: (a) PV generation forecast and (b) BiLSTM disaggregation results.

in the reported results. Among the Regression, KNN, and SVM models, which are ML rather than DL techniques, the Regression and SVM algorithms deliver the best *RMSE*, due to KNN's reliance on a single hyper-parameter, i.e., the number of nearest neighbors, which is inadequate for accurately modeling the complex input-output relationships in predictive tasks. Furthermore, KNN is prone to overfitting due to its limited hyper-parameters and lack of regularization mechanisms.

Observe in Table V that all models closely approximate the estimated PV output for 2021. However, the proposed BiLSTM DL model consistently provides accurate estimations across various weather conditions, underscoring its enhanced performance relative to existing methods.

The CNN-LSTM model exhibits superior capability for identifying local patterns and non-linear correlations between various features and electrical load compared to the LSTM model. This results in lower forecasting error for the CNN-LSTM method, which achieves better alignment between the estimated and predicted PV generation, as illustrated in Table V and Fig. 16 to Fig. 18.

To better compare the performance of the ML and DL models, a rolling time window analysis was applied to the obtained metrics to assess the differences between the algorithms over time. Thus, Fig. 19 to Fig. 21 display the performance measurements, specifically MAE and RMSE, for both ML and DL algorithms, with lower RMSEs and MAEs indicating better predictive accuracy and model results. For a fair evaluation, time samples when there is no PV generation were excluded to prevent skewing the results. Observe that, for all models, error values increase during peak PV generation at around noon, which suggests that the highest prediction error are likely due to rapid fluctuations in irradiance and temperature variations at peak solar generation. Note that the most significant prediction error differences between the ML and DL algorithms occur in May, which corresponds to high solar activity and variability, while lower error differences are observed in December, which has the lowest peak PV power generation due to reduced sunlight.

TABLE V
QUANTITATIVE COMPARISON OF ML/DL MODELS FOR PV GENERATION FORECASTING

Period	Error	Regression	KNN	SVM	LSTM	CNN-LSTM	BiLSTM
May, 2021	$MAE[MW]$	0.17	0.38	0.25	0.26	0.18	0.13
	$RMSE[MW]$	0.27	0.47	0.32	0.36	0.27	0.23
	$R^2[p.u.]$	0.95	0.87	0.94	0.93	0.97	0.97
	$SD[MW]$	1.33	1.33	1.33	1.40	1.66	1.67
	$PA[p.u.]$	0.97	0.94	0.97	0.96	0.98	0.99
July, 2021	$MAE[MW]$	0.10	0.35	0.18	0.16	0.16	0.09
	$RMSE[MW]$	0.17	0.42	0.24	0.23	0.21	0.16
	$R^2[p.u.]$	0.98	0.89	0.96	0.96	0.98	0.96
	$SD[MW]$	1.28	1.28	1.28	1.31	1.62	1.28
	$PA[p.u.]$	0.99	0.97	0.98	0.98	0.99	0.98
Dec, 2021	$MAE[MW]$	0.09	0.16	0.11	0.06	0.09	0.01
	$RMSE[MW]$	0.16	0.28	0.17	0.12	0.16	0.12
	$R^2[p.u.]$	0.96	0.89	0.95	0.98	0.96	0.98
	$SD[MW]$	0.88	0.88	0.88	0.87	0.88	0.88
	$PA[p.u.]$	0.98	0.95	0.98	0.99	0.98	0.99

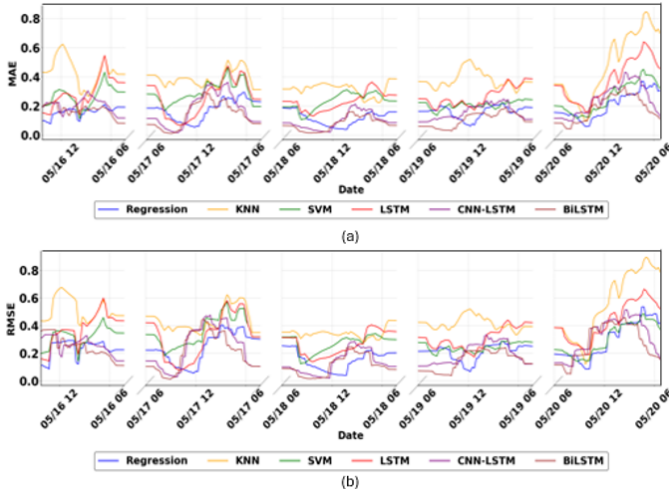


Fig. 19. Rolling window performance metric analysis for different algorithms over the specified date range in May 2021: (a) MAE and (b) RMSE.

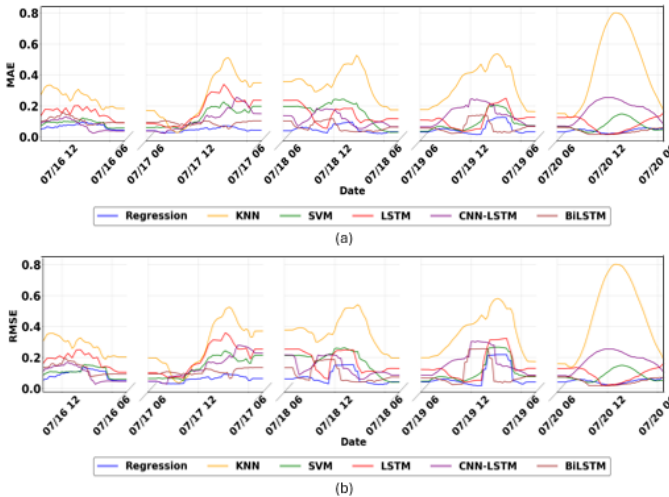


Fig. 20. Rolling window performance metric analysis for different algorithms over the specified date range in July 2021: (a) MAE and (b) RMSE.

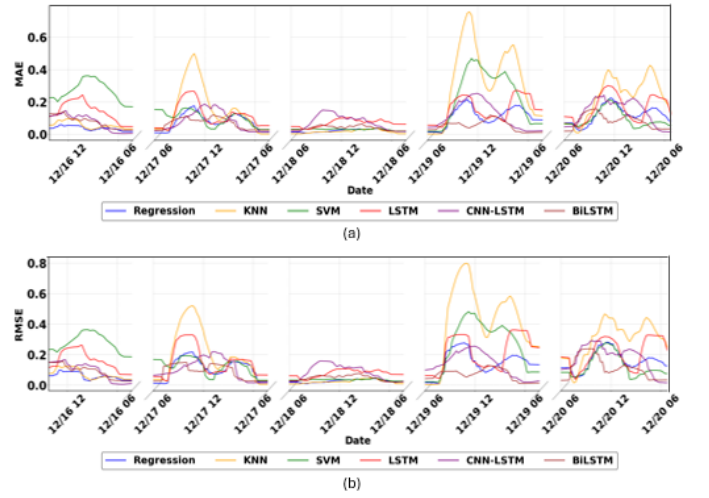


Fig. 21. Rolling window performance metric analysis for different algorithms over the specified date range of December 2021: (a) MAE and (b) RMSE.

Among the ML models, observe that KNN exhibits the highest error in Fig. 19 to Fig. 21, highlighting its limitations in capturing complex temporal dependencies. In contrast, the Regression model demonstrates better performance, followed by SVM, as both benefit from their ability to map non-linear relationships more effectively. On the other hand, note that the DL models, particularly BiLSTM and CNN-LSTM, consistently show lower RMSE and MAE values, thus indicating superior performance. The BiLSTM model yields the most accurate predictions, with its values closely matching actual observations. This can be attributed to its bidirectional processing capability, which allows it to learn patterns from both past and future time steps, enhancing predictive accuracy. Similarly, CNN-LSTM leverages convolutional layers to extract spatial-temporal features before passing them through its LSTM layers, improving its ability to generalize across different time periods.

The various proposed identification techniques can be further compared using the RMSE and MAE distributions illustrated in the boxplots depicted in Fig. 22 and Fig. 23, which allow comparing the medians and variability for the different months previously discussed, providing details of

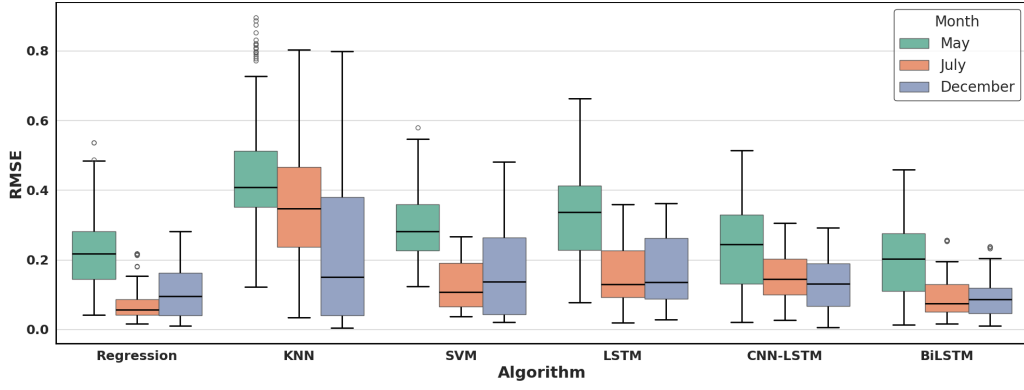


Fig. 22. RMSE boxplots across models and months.

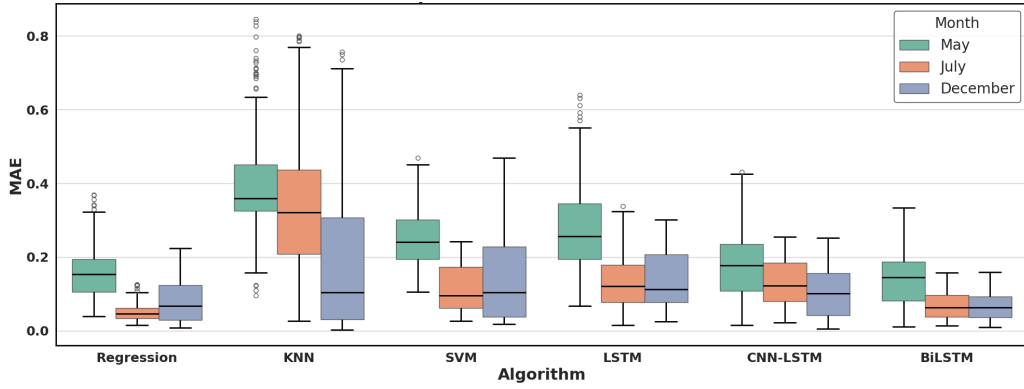


Fig. 23. MAE boxplots across models and months.

the data's distribution characteristics for different seasonal conditions. The boxes represent the range between the first (25% percentile) and third (75% percentile) data quartiles, while the whiskers show the range between minimum and maximum data points, with the open circles reflecting data outliers beyond 1.5 times the shown Interquartile Ranges (IQRs). Observe that the median RMSE and MAE values for all models are the highest in May, due to increased solar power generation and greater variability during this period. The IQRs are not consistent across months for each algorithm, with May generally showing the widest IQRs, especially for the Regression, LSTM, CNN-LSTM, and BiLSTM models, indicating higher dispersion and sensitivity during peak solar generation. In contrast, July tends to have the lowest overall spread due to the highest radiation days and temperatures of the year, as expected. Note that for the KNN and SVM models, December shows broader IQRs, with KNN displaying the highest IQRs among all algorithms and months, thus reflecting its higher prediction errors previously highlighted.

The demonstrated superiority of DL algorithms over ML techniques confirms the effectiveness of incorporating temporal feature extraction and sequential learning mechanisms. By leveraging memory-based architectures, DL models can better capture long-term dependencies, making them more resilient to variations in PV generation patterns. These results highlight the advantage of using DL approaches for improved solar power prediction.

V. CONCLUSIONS

This paper introduced ML and DL based approaches for disaggregating PV power generation from net-load measurements, specifically separating the contributions of small PV systems from actual demand load. The proposed approach considered a range of data inputs, including net-load measurements, meteorological and irradiance data, satellite imagery, and temporal information. A Mask RCNN model was first employed to detect solar PV panels from high-resolution satellite images, enabling accurate estimation of solar capacity and PV power generation. The results for the case study in the Waterloo region in Ontario, Canada, demonstrated that the identified PV capacity achieved an accuracy of approximately 98%. The synthetic PV data generated using a reliable software tool was then used for effective net-load disaggregation. The BiLSTM algorithm was shown to have superior precision with respect to estimated PV generation compared to several other ML/DL approaches. The results show that the proposed method should be useful for various feeder applications, offering robust performance for practical use.

While this paper focused solely on PV systems, other DERs such as heat pumps (HPs) and EVs pose significant challenges in modern power system operation. Thus, future research on net-load disaggregation will aim to address the identification of HPs, given their rapid integration in power grids and the expected significant demand impact.

ACKNOWLEDGMENT

The feedback and relevant information provided by Allen Chan from Enova Power Corp. for the Waterloo region in Ontario, Canada, is gratefully acknowledged.

REFERENCES

- [1] F. Wang, K. Li, X. Wang, L. Jiang, J. Ren, Z. Mi, M. Shafie-khah, and J. P. Catalão, "A distributed PV system capacity estimation approach based on Support Vector Machine with customer net load curve features," *Energies*, vol. 11, no. 7, p. 1750, 2018.
- [2] A. O. Chen Xinlei, "Data efficient solar disaggregation with behind-the-meter energy resources," *Sustainable Energy, Grids and Networks*, vol. 32, p. 100813, 2022.
- [3] E. C. Kara, C. M. Roberts, M. Tabone, L. Alvarez, D. S. Callaway, and E. M. Stewart, "Towards real-time estimation of solar generation from micro-synchrophasor measurements," *arXiv:1607.02919*, 2016.
- [4] J. M. Malof, K. Bradbury, L. M. Collins, and R. G. Newell, "Automatic detection of solar photovoltaic arrays in high resolution aerial imagery," *Applied Energy*, vol. 183, pp. 229–240, 2016.
- [5] J. de Hoog, S. Maetschke, P. Ilfrich, and R. R. Kolluri, "Using satellite and aerial imagery for identification of solar PV: State of the art and research opportunities," in *ACM International Conference on Future Energy Systems*, 2020, pp. 308–313.
- [6] W. Kanchana and J. G. Singh, "Detection of invisible solar PV installations in power distribution system: A review," in *IEEE International Conference and Utility Exhibition on Energy, Environment and Climate Change (ICUE)*, 2022, pp. 1–8.
- [7] B. Bayer, P. Matschoss, H. Thomas, and A. Marian, "The German experience with integrating photovoltaic systems into the low-voltage grids," *Renewable Energy*, vol. 119, pp. 129–141, 2018.
- [8] E. Vrettos, E. Kara, E. Stewart, and C. Roberts, "Estimating PV power from aggregate power measurements within the distribution grid," *Journal of Renewable and Sustainable Energy*, vol. 11, no. 2, 2019.
- [9] A. F. M. Jaramillo, D. M. Laverty, D. J. Morrow, J. M. del Rincon, and A. M. Foley, "Load modelling and non-intrusive load monitoring to integrate distributed energy resources in low and medium voltage networks," *Renewable Energy*, vol. 179, pp. 445–466, 2021.
- [10] S. Bahramara, A. Mazza, G. Chicco, M. Shafie-khah, and J. P. Catalão, "Comprehensive review on the decision-making frameworks referring to the distribution network operation problem in the presence of distributed energy resources and microgrids," *International Journal of Electrical Power & Energy Systems*, vol. 115, p. 105466, 2020.
- [11] N. Batra, R. Kukunuri, A. Pandey, R. Malakar, R. Kumar, O. Krystallakos, M. Zhong, P. Meira, and O. Parson, "Towards reproducible state-of-the-art energy disaggregation," in *ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, 2019, pp. 193–202.
- [12] "Multi-task deep learning for large-scale building detail extraction from high-resolution satellite imagery."
- [13] P. Gotseff, J. Cale, M. Baggu, D. Narang, and K. Carroll, "Accurate power prediction of spatially distributed PV systems using localized irradiance measurements," in *IEEE PES General Meeting*, 2014, pp. 1–5.
- [14] J. Alonso-Montesinos, F. Batlles, and J. Bosch, "Beam, diffuse and global solar irradiance estimation with satellite imagery," *Energy Conversion and Management*, vol. 105, pp. 1205–1212, 2015.
- [15] A. Sezer and A. Altan, "Detection of solder paste defects with an optimization-based deep learning model using image processing techniques," *Soldering & Surface Mount Technology*, vol. 33, no. 5, pp. 291–298, 2021.
- [16] G. Jiménez and D. Racocanu, "Deep learning for semantic segmentation vs. classification in computational pathology: application to mitosis analysis in breast cancer grading," *Frontiers in Bioengineering and Biotechnology*, vol. 7, p. 145, 2019.
- [17] A. Altan, S. Karasu, and E. Zio, "A new hybrid model for wind speed forecasting combining long short-term memory neural network, decomposition methods and Grey Wolf optimizer," *Applied Soft Computing*, vol. 100, p. 106996, 2021.
- [18] E. G. Vera, C. A. Cañizares, M. Pirnia, T. P. Guedes, and J. D. M. Trujillo, "Two-stage stochastic optimization model for multi-microgrid planning," *IEEE Transactions on Smart Grid*, vol. 14, no. 3, pp. 1723–1735, 2022.
- [19] S. Lee, S. Iyengar, M. Feng, P. Shenoy, and S. Maji, "Deeproof: A data-driven approach for solar potential estimation using rooftop imagery," in *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 2105–2113.
- [20] J. Yu, Z. Wang, A. Majumdar, and R. Rajagopal, "Deepsolar: A machine learning framework to efficiently construct a solar deployment database in the United States," *Joule*, vol. 2, no. 12, pp. 2605–2617, 2018.
- [21] J. M. Malof, L. M. Collins, and K. Bradbury, "A deep convolutional neural network, with pre-training, for solar photovoltaic array detection in aerial imagery," in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2017, pp. 874–877.
- [22] Z. Huang, T. Mendis, and S. Xu, "Urban solar utilization potential mapping via deep learning technology: A case study of Wuhan, China," *Applied Energy*, vol. 250, pp. 283–291, 2019.
- [23] R. Castello, S. Roquette, M. Esguerra, A. Guerra, and J.-L. Scartezzini, "Deep learning in the built environment: Automatic detection of rooftop solar panels using Convolutional Neural Networks," in *Journal of Physics: Conference Series*, vol. 1343, no. 1, 2019, p. 012034.
- [24] P. Li, H. Zhang, Z. Guo, S. Lyu, J. Chen, W. Li, X. Song, R. Shibasaki, and J. Yan, "Understanding rooftop PV panel semantic segmentation of satellite and aerial images for better using machine learning," *Advances in Applied Energy*, vol. 4, p. 100057, 2021.
- [25] T. Zhong, Z. Zhang, M. Chen, K. Zhang, Z. Zhou, R. Zhu, Y. Wang, G. Lü, and J. Yan, "A city-scale estimation of rooftop solar photovoltaic potential based on deep learning," *Applied Energy*, vol. 298, p. 117132, 2021.
- [26] Q. Chen, F. Wang, B.-M. Hodge, J. Zhang, Z. Li, M. Shafie-Khah, and J. P. Catalão, "Dynamic price vector formation model-based automatic demand response strategy for PV-assisted EV charging stations," *IEEE Transactions on Smart Grid*, vol. 8, no. 6, pp. 2903–2915, 2017.
- [27] F. Wang, K. Li, C. Liu, Z. Mi, M. Shafie-Khah, and J. P. Catalão, "Synchronous pattern matching principle-based residential demand response baseline estimation: Mechanism analysis and approach description," *IEEE Transactions on Smart Grid*, vol. 9, no. 6, pp. 6972–6985, 2018.
- [28] L. Massidda and M. Marrocu, "Decoupling weather influence from user habits for an optimal electric load forecast system," *Energies*, vol. 10, no. 12, p. 2171, 2017.
- [29] Y. Yu, X. Wen, J. Zhao, Z. Xu, and J. Li, "Co-planning of demand response and distributed generators in an active distribution network," *Energies*, vol. 11, no. 2, p. 354, 2018.
- [30] L. Waswa, M. J. Chihota, and B. Bekker, "A probabilistic estimation of PV capacity in distribution networks from aggregated net-load data," *IEEE Access*, vol. 9, pp. 140 358–140 371, 2021.
- [31] Esri, "ArcGIS pretrained models: Introduction to the model," [Online]. Available: <https://doc.arcgis.com/en/pretrained-models/latest/imagery/introduction-to-solar-panel-detection-usa.htm>, 2024, Accessed: Aug. 28, 2024.
- [32] X.-Y. Zhang, S. Kuenzel, P. Guo, L. Yin, and C. Watkins, "A hybrid data-driven online solar energy disaggregation system from the grid supply point," *Complex & Intelligent Systems*, vol. 9, no. 4, pp. 3695–3716, 2023.
- [33] Independent Electricity System Operator (IESO), "Supply overview: Transmission-connected generation," [Online]. Available: <https://www.ieso.ca/en/Power-Data/Supply-Overview/Transmission-Connected-Generation>, 2024, Accessed: Aug. 28, 2024.
- [34] I. Goodfellow, *Deep Learning*. MIT Press, 2016.
- [35] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep learning for computer vision: A brief review," *Computational Intelligence and Neuroscience*, vol. 2018, no. 1, p. 7068349, 2018.
- [36] H. Li, B. Herfort, and A. Zipf, "Estimating openstreetmap missing built-up areas using pre-trained deep neural networks," in *GILE Conference on Geographic Information Science*, 2019, pp. 17–20.
- [37] National Renewable Energy Laboratory (NREL), "NREL system advisor model (SAM)," [Online]. Available: <https://sam.nrel.gov/weather-data/weather-data-publications.html>, 2024, Accessed: Sep. 11, 2024.
- [38] P. Anandhi and E. Nathiya, "Application of linear regression with their advantages, disadvantages, assumption and limitations," *International Journal of Statistics and Applied Mathematics*, vol. 8, pp. 133–137, 2023.
- [39] "Electrical load forecasting: A deep learning approach based on K-nearest neighbors."
- [40] "Distance and similarity measures effect on the performance of K-nearest neighbor classifier—a review."
- [41] "PV forecasting using support vector machine learning in a big data analytics context."
- [42] "Exploring kernel machines and support vector machines: Principles, techniques, and future directions."

- [43] "Solar power forecasting using cnn-lstm hybrid model."
- [44] U. Javed, K. Ijaz, M. Jawad, I. Khosa, E. A. Ansari, K. S. Zaidi, M. N. Rafiq, and N. Shabbir, "A novel short receptive field based dilated causal convolutional network integrated with Bidirectional LSTM for short-term load forecasting," *Expert Systems with Applications*, vol. 205, p. 117689, 2022.
- [45] "Short-term traffic flow prediction for urban road sections based on time series analysis and LSTM_BiLSTM method."
- [46] National Renewable Energy Laboratory (NREL), "NREL system advisor model (SAM)," [Online]. Available: <https://sam.nrel.gov/download>, 2024, Accessed: Sep. 11, 2024.
- [47] C. Seger, "An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing," [Online]. Available: <https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1259073&dsid=-6062>, 2018, Accessed: Sep. 11, 2024.
- [48] A. Géron, *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media, 2022.
- [49] "The pascal visual object classes (voc) challenge."
- [50] "Toward global rooftop pv detection with deep active learning."
- [51] "Semantic segmentation of clouds in satellite imagery using deep pre-trained u-nets."
- [52] "A novel deep learning-based single shot multibox detector model for object detection in optical remote sensing images."
- [53] Z. Hameed and B. Garcia-Zapirain, "Sentiment classification using a single-layered BiLSTM model," *IEEE Access*, vol. 8, pp. 73 992–74 001, 2020.
- [54] D. P. Kingma, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [55] "Forecasting PM2.5 concentration using a single-dense layer BiLSTM method."
- [56] C. Guo, G. Liu, and C.-H. Chen, "Air pollution concentration forecast method based on the deep ensemble neural network," *Wireless Communications and Mobile Computing*, vol. 2020, no. 1, p. 8854649, 2020.
- [57] *Pattern recognition and machine learning*.
- [58] O. Tarkhaneh and H. Shen, "Training of feedforward Neural Networks for data classification using hybrid particle Swarm optimization, Mantegna Lévy flight and neighborhood search," *Heliyon*, vol. 5, no. 4, 2019.



Hediyyeh Safari (Graduate Student Member, IEEE) is currently pursuing the Ph.D. degree in Electrical and Computer Engineering at the University of Waterloo, ON, Canada. Her research focuses on renewable energy integration, forecasting and load disaggregation, synthetic data generation, statistical analysis, and the application of AI and machine learning in smart grids. She has collaborated with the Independent Electricity System Operator (IESO), Ontario, on projects related to distributed photovoltaic (PV) system visibility and heat pump (HP)

power disaggregation. Her interests also include data-driven modeling, satellite imagery analytics, and enhancing situational awareness in transmission and distribution networks.



Dr. Claudio Cañizares (Fellow, IEEE) is a University Professor, Hydro One Chair, and Executive Director of WISE at Waterloo, where he has been since 1993. His work on modeling, simulation, computation, stability, control, and optimization of power and energy systems is highly cited and recognized. He is the past IEEE Trans. Smart Grid EIC and an IEEE and PES Boards' Director, and is a Fellow of the IEEE, Royal Society of Canada, Canadian Academy of Engineering, and Chinese Society for Electrical Engineering. He has received the IEEE PES 2025

Ramakumar Family Renewable Energy Excellence and 2017 Outstanding Educator Awards, the 2016 IEEE Canada Electric Power Medal, and various other awards, recognitions, and leadership appointments from IEEE, PES, Waterloo, and Chinese universities.



Engineering specializing in Energy Systems from the University of Toronto in 2018. He is a professional engineer in Ontario, Canada.

Daniel Sohm (Member, IEEE) has 14 years of experience at Independent Electricity System Operator validating equipment performance, simulating events on the power system, managing ancillary service contracts, integrating energy storage into the wholesale electricity market, performing research relating to energy storage and ancillary services, and in transmission planning. He received a B.A.Sc. degree in Electrical and Computer Engineering from the University of Waterloo in 2011 and a Master of Engineering degree in Electrical and Computer



Elyas Ahmed, P.Eng (Member, IEEE) is a Senior Engineer Advanced Technology Research in the Transmission Planning Projects team at the IESO in Ontario. His work spans energy storage technologies, system modelling and optimization. He holds a bachelor's degree in Electronics Engineering from Simon Fraser University and a graduate degree in Management Analytics from the Smith School of Business at Queen's University.



Dr. Ismael El-Samahy (Member, IEEE) received his Ph.D. degree in Electrical and Computer Engineering from the University of Waterloo, Canada, with a focus on power system operations and economics. He has over 17 years of experience in the energy sector in Ontario, including leadership roles in transmission planning, electricity markets, and system operations. Ismael is currently the Senior Manager of Transmission Planning Projects at the Independent Electricity System Operator (IESO) in Ontario, Canada. Dr. El-Samahy is also a member of both the Advisory Board of the Canadian University of Dubai in the UAE, and the Industry Advisory Board for the Department of Electrical Engineering and Computer Science at York University in Ontario, Canada. His active engagement with both academia and industry reflects a deep commitment to advancing power system engineering through applied innovation, thought leadership, and collaborative development.



communities.

Maurice Dusseault is an Emeritus Professor at the University of Waterloo focusing on research in subsurface engineering. This domain includes energy storage as fuel (H₂, CH₄), heat, or compressed air (or CO₂). Also, he continues to work on aspects of deep waste disposal of fluids and slurried wastes, including biosolids injection, heavy metal contaminated soils slurification, and CO₂ sequestration. Finally, he is involved with geomechanics aspects of geothermal systems integrated with heat pumps, heat storage and some power generation for remote