International Journal of Intelligent Information Technologies
Volume 19 • Issue 1

Anomaly Detection in Renewable Energy Big Data Using Deep Learning

Suzan MohammadAli Katamoura, King Saud University, Saudi Arabia*
https://orcid.org/0000-0003-4758-2143

Mehmet Sabih Aksoy, King Saud University, Saudi Arabia

ABSTRACT

This work aims to review the literature on anomaly detection (AD) in renewable energy. Due to the significance of the RE data quality and sensor performance, it is crucial to ensure that the measurement device works correctly and maintains data accuracy. The review identifies the relevant studies on big data anomaly detection in the energy field and synthesizes the related techniques. Also, the study shows a need for segmentation annotations for solar system electroluminescence imagery complicating the domain development of anomaly segmentation approaches. Consequently, most processes create machine learning (ML) models using semi-supervised techniques. Still, these approaches need more generalization regarding variation in environmental or systematic conditions. Furthermore, the studies discussed here focus on existing algorithms that used big data and AD to propose an improved analysis framework. Finally, the work presents a framework to solve the problem of identifying sensors' issues that will appear in data anomalies.

KEYWORDS

Anomaly Detection, Big Data, Deep Learning, Renewable Energy, Solar Data

INTRODUCTION

Renewable Energy (RE) attracts countries worldwide for many potentials and benefits. Thus, Saudi Arabia has increased its focus on including renewable resources in the national energy mix. It is currently developing and investing in many energy systems. The primary reasons are contributing to its climate obligations and diversifying its economy from being fossil fuel-dependent (Barhoumi et al., 2020).

A significant challenge for the global energy supply is the large integration of RE sources, such as solar energy, into existing or future energy supply structures. An electrical operator must maintain a precise balance of electricity production and consumption. Additionally, the operator frequently needs help maintaining this balance with conventional and controllable energy production systems, particularly in small or non-interconnected electrical grids like islands. Moreover, the electrical system's reliability depends on its ability to maintain quality and continuity of service to customers.

DOI: 10.4018/IJIIT.331595 *Corresponding Author

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

Then, the complexity of the system management across multiple time horizons, considering RE's erratic behavior, is increased (Espinar et al., 2010).

Solar production's intermittent and uncontrollable characteristics cause several other issues, including voltage fluctuations, local power quality, and stability issues. So, predicting solar system output power is necessary for the smooth functioning and the best control of energy fluxes into the solar system, estimating reserves, scheduling the power system, managing congestion, and reducing electricity production costs (Voyant et al., 2017).

Because of the significant increase in solar power generation, forecasting solar yields is becoming increasingly important to avoid significant variations in renewable electricity production. Various systems are being developed to control fluctuations and maintain power quality continuity through forecasting horizons that range from 5 minutes to several days. Moreover, several studies have outlined the reasons to forecast solar radiation for various solar systems. Furthermore, the predicted data's time step will vary depending on the objectives and forecasting horizon (Yang et al., 2020).

In particular, solar systems have the potential to reduce the reliance on carbon-intensive energy sources significantly. Hence, progress over the last 60 years has been made to improve their efficiency. However, due to an anomaly in the output data, current efficiency levels are relatively low. In addition, various abnormalities can occur, preventing solar systems from operating at full capacity. As a result, detecting and repairing these issues is critical to ensure maximum efficiency (Blaga et al., 2019).

In Saudi Arabia, King Abdullah City for Atomic and Renewable Energy (K.A.CARE) collects solar irradiances and the associated meteorological data from 46 ground-based stations across the country to help in the decision-making process related to the country's projects. The solar irradiance data are collected using different devices that measure incidents on a sensor surface. However, in a hostile climate such as Saudi Arabia, the devices' performance is affected by various elements, such as dust particles accumulation on sensors (Zell et al., 2015) and missing data due to equipment failure, sensor cleaning procedure, or calibration (Vinisha & Sujihelen, 2022). As a result, this can influence data quality severely due to the different anomaly sources.

Since applications depend principally on energy data, such as forecasting energy production or consumption, they necessitated using effective models and accurate input data to give a precise output (prediction). However, with the vast time-series RE data, i.e., solar radiation measurements, it is difficult to assure its correctness. Thus, the most critical task is to confirm that the data collected from different instruments under different weather conditions is accurate. Accurate data will reflect on an accurate model. Thus, the question is how to confirm the data's correctness.

Therefore, this work reviews the literature to understand and analyze the various prediction and anomaly-detection methodologies for big data, i.e., RE or solar data. Furthermore, the work aims to identify the gaps and propose a recommendation for improving the existing techniques to detect data correctness.

Many studies are available, but finding one dedicated to a single technique is rare. Thus, the paper discusses several methodologies used to predict anomalies in solar radiation and the parameters used to estimate model performances to elaborate on the development of anomaly detection (AD) methods, including deep learning (DL).

The present paper contributes to:

1. Identify the relevant studies on big data AD in the energy field.
2. Synthesize the related concepts or techniques such as supervised, unsupervised, and semi-supervised tasks.
3. Recommend a framework that can be employed successfully to solve the problem of validating solar data correctness concerning the issues that might affect data accuracy due to sensor issues.

The remaining part of the paper is divided into the following: section 2 gives the necessary background about the linked concepts, section 3 reviews the related works, the corresponding findings

are discussed in section 4, the proposed AD framework is introduced in section 5, and section 6 concludes the review paper.

BACKGROUND

Renewable Energy and Solar Data

Since the electricity demand in Saudi Arabia has increased, with an annual growth rate of about 6%, several solar photovoltaic (PV) projects were announced to be conducted over the next decade. These projects are planned with an investment of more than 33 billion Saudi Riyals to cover about 32% of the current peak power consumption of 61.743 GW by supplying electricity with very competitive costs. So, Saudi Arabia will target 27.3 GW of RE by 2023, most of which 20 GW will be solar PV, while wind and concentrated solar power (CSP) will be 7.3 GW. Then, by 2030, Saudi Arabia aims to expand RE projects up to 58.7 GW, with significant measures taken by the government following National Vision 2030 (AlOtaibi et al., 2020). Therefore, this will increase demand for the country's RE industry, particularly solar energy (Administration, 2021).

Additionally, solar radiation is an essential energy source on earth because of its significant role in surface radiation balance and weather and climate extremes. Accurate solar radiation prediction is also critical in the solar industry and climate research (Iyengar et al., 2018). Solar energy's rapid industrial growth increases interest in RE from smart grids and plants. However, detecting anomalies in solar radiation systems is difficult (Hu, 2012).

The country has invested considerably in installing ground-based network stations to collect solar irradiance data in the last ten years. Mainly, they measure Direct Normal Irradiance (DNI), Global Horizontal Irradiance (GHI), and Diffuse Horizontal Irradiance (DHI) over several locations with different climatic conditions (along with other weather data). The measurements were collected since 2013 at a 1-min resolution. Such measurements are crucial in assisting government and investors in decision-making regarding the project planning and development of large solar power plants. This will help increase the share of renewable sources in the energy mix and achieve the 2030 renewable targets. These collected data are processed and uploaded to the Renewable Resources Atlas portal by a dedicated team and managed by K.A.CARE (Zell et al., 2015).

K.A.CARE has no single system to validate the correctness of the collected data nor an automated warning feature that notifies the operators to take action. Instead, the team of engineers runs a quality check process on 1-minute data resolution, developed by the National Renewable Energy Laboratory (NREL), to investigate data issues and take corrective actions directly or by reporting the issue to the maintenance team. The quality assurance process consists of many steps that involve checking errors logically (built into the algorithm) as data is captured and stored at the datacenter, and manually through daily data checks station operations (through data plots) adhere to expected behaviors, limits, and scheduled maintenance. Also, the process includes running a series of automated and semi-automated software tools at different frequencies (weekly or monthly). These software tools include the SERIQC program to flag data issues based on an acceptable range of values and Ratios software that plots the average readings. Additionally, the team performs manual analysis of results to confirm data is correct and accurate. However, these steps are disconnected, require excessive human interference, and depend on an individual's experience.

The steps require a quality team member to conduct them following a specific order with no direct interaction between the different tools or steps, i.e., each step is isolated and can be completed by a different agent. Nevertheless, the output of some steps can be the input for another step and can affect the result remarkably. Also, since the knowledge of each member differs based on the experience and exposure to different scenarios and familiarity with used programs, human errors are considered the main factor of accuracy issues. Furthermore, the delay in discovering issues and notifying the operators may cause measurement loss or inaccurate values. Consequently, the data will

suffer from incompleteness (between 6-35%) and inaccuracy (up to 9%), which affects its correctness and subsequent decisions and economic impact.

Big Data

Big data describes datasets with a size and type that traditional relational databases cannot capture, manage, or process. Also, big data characterizes high volume, velocity, and variety. To handle big data complexity, some techniques or tools like Artificial intelligence (AI), mobile, and the Internet of Things (IoT) are used to drive data complexity through new forms of data. An illustration of big data is a dataset generated in real-time at an enormous scale, such as those from sensors, devices, video/audio, and networks (Nazarov et al., 2022).

Accordingly, we can conclude two essential facts. First, RE data is a big time-series of massive, diverse data collected instantly. For instance, solar data consists of time series measurements, related meteorological data, and images, making it challenging to judge data correctness. Similarly, (Ghofrani & Alolayan, 2018) indicated that solar PV power is a function of solar radiation exhibiting unique features that complicate their prediction and make solar power forecasting challenging. Second, due to planned investments and goals, accurate solar forecasting, hence solar data, enhances the value of RE by improving these resources' reliability and economic feasibility. Consequently, data anomalies in solar measurements can significantly affect RE (solar) projects' performance and investment.

To further emphasize the previous conclusion, (Nazarov et al., 2022) in his work indicated that implementing big data technologies into energy systems improves energy data quality. Consequently, the energy industry, including production processes, transmission, and sale of electrical energy, are improved too. Moreover, making optimal decisions aimed at modernizing the energy sector is supported.

Different anomaly sources can influence data quality severely and can exhibit data abnormalities. For example, RE data accuracy is highly dependent on weather conditions, which include numerous variables (e.g., dust particles accumulation on sensors) (Zell et al., 2015). Additionally, equipment failure that causes missing data, device misalignment, and sensor cleaning procedure or calibration can introduce data variation (Vinisha & Sujihelen, 2022). This highlighted a vital data issue with real-time energy generation forecasting. To solve this problem, big data techniques such as machine learning (ML), data mining (DM), and DL are used and discussed in previous studies. Various DM methods analyze big data, while machine learning and DL algorithms implement forecasting models and AD methods (Hu, 2012); (Moritz Benninger, 2019); (Natarajan et al., 2020).

Anomaly Detection

DM has improved computer system performance and functionality, and AD is a prominent data mining algorithm and an important research topic. AD is a process for detecting system or data abnormalities diagnosing and resolving system issues (Harimi & Shayegan Fard, 2020). Also, AD in solar systems can be accomplished using various techniques, ranging from traditional statistics to DM and DL. Although such methods are well suited to detecting anomalies in solar systems, they typically necessitate data that is only sometimes readily available for AD. Several methods for AD are based on data from the solar system and environment, such as solar irradiation and temperature (Branco et al., 2020).

It is challenging to detect anomalies in massive volume time series data using standard methods. This is because the standard approaches divide large amounts of data into several small samples, which are subsequently analyzed. However, doing so increases the complexity of segmentation algorithms and makes controlling the risk of data segmentation harder. Therefore, to solve this problem, multiple machine learning techniques are required, such as regression, decision trees, and Support Vector Machine (SVM), along with DL, which offers better performance in training time and accuracy (Soni, 2021).

Deep Learning

Deep learning (DL) is an emerging technology dealing with big data, a computer science subfield classified as an AI method. It is concerned with the design and study of systems learning from data sets without explicit programming. Additionally, DL has various tools and algorithms, proving robustness, accuracy, and many advantages. Moreover, DL techniques can implement some optimizers to enhance the detection rate and performance (Sarker, 2021).

DL can be used in diverse domains, e.g., RE. DL models discover relationships between inputs and outputs even when representation is impossible; this property allows DL models to be used in a wide range of applications, including classification, DM, and forecasting. Literature examines DL and deterministic methods for solar forecasting (Roka et al., 2022). For example, classification and DM are intriguing in RE data forecasting because they work well with large datasets, while DL models can handle preprocessing and data preparation. Also, learning performance increases as the dataset increases (Sarker, 2021). Thus, DL models can be applied to forecasting or AD problems, i.e., models for GHI forecasting based on other meteorological and geographical parameters, including time-series models that only consider historically observed solar irradiance data as input features and hybrid models that consider both solar irradiance and other variables as exogenous variables (Roka et al., 2022).

In this regard, using the most recent advancements in ML technology to accurately and timely reveal various energy system abnormalities is critical. Additionally, AD in modern power plants reduces downtime and increases efficiency (Buerhop-Lutz et al., 2018). So, data examination is the best way to detect the system anomaly. However, that requires a robust model or technique to identify data variations from different sources.

LITERATURE REVIEW

Several studies have been conducted to investigate AD techniques in big data and energy systems. These techniques cover various DM algorithms, including regression, support vector machines, neural networks, k-nearest-neighbors classification, clustering, and DL. The literature review attempts to group studies thematically when possible and then organizes them within each group chronologically to show the progress of anomaly prediction methods.

General Surveys

Various AD methods were surveyed, including classification, nearest neighbor, clustering, statistical, spectral, information-theoretic, and graphing techniques. The study concluded that selecting the best AD method for a specific problem depends on criteria such as input data, kind of anomalies, output data, and domain expertise (Toshniwal et al., 2020).

(Hu, 2012) compared several supervised and unsupervised algorithms (for detecting and classifying sensor abnormalities, including the auto-regressive integrated moving average model (ARIMA), SVM, NNs, and k-nn classification. The work proposed model performs in real-life situations considering ten types of anomaly sources. The result demonstrated that using the k-nn classification for AD correctly labels 97% of the testing dataset anomalies (Hu, 2012). However, the algorithms perform well with small test cases and address univariate time series data, while the targeted data are multivariate.

In (Himeur et al., 2021) work, they systematically surveyed AI-based AD systems for construction energy consumption. The authors discussed various supervised and unsupervised approaches, including clustering, one-class classification, dimensionality reduction NNs, regression, probabilistic models, and traditional classification. Also, they explained feature extraction strategies using distance-based, time-series analysis, density-based, graph-based, and compressive sensing concepts. Additionally, the authors found that no AD scheme could discover all anomalies using low-dimensional subspaces

due to the data complexity and other variables affecting power utilization. Thus, they introduced frameworks adopting semi-supervised AD methods, such as boosting and bagging, to identify normal power usage rather than labeling anomalous patterns. However, the research needs further work with industry to consider scalability, easy implementation, and privacy preservation in developing power AD systems. Besides, the proposed unsupervised learning models can only detect one type of anomaly.

The main conclusion of these studies was that the choice of DM techniques for AD depends on several factors, primarily the data and problem description.

Supervised Algorithms

Detecting the malfunction of the solar sensor using supervised algorithms was covered thoroughly in the literature. In general, the performance of the supervised algorithms is satisfactory. Among all, an SVM deals with high-dimensional data with efficient memory usage, handles nonlinear features, and does not make a presumption about the underlying data distribution.

(Moritz Benninger, 2019) This work presented a technique for monitoring systems data by detecting failures using KNN and a 1-class SVM (OCSVM). The learning algorithms considerably reduce the measuring effort and offer reliable anomaly monitoring. However, this work has specific hardware requirements.

(Harrou et al., 2019) They proposed a model-based AD technique for inspecting the stable portion of the sensors and momentary shading using a model to outline the ordinary nature of the supervised sensor system and to form residuals for fault detection. Then, the 1-SVM fault prediction model is applied to residuals to assess the dissimilarity between normal and abnormal characteristics. The proposed technique outperforms unsupervised conventional binary clustering algorithms (i.e., K-means, mean-shift, expectation–maximization, Birch, and agglomerative). However, the model addressed only one type of anomaly and performed well with specific PV sensors but was unsuitable for the solar sensors.

Natarajan et al., 2020 proposed a new technique for device fault detection based on thermal image processing with an SVM tool that classified the attributes as defective or non-defective. The technique extracts the device's operating features and compares them to the standard features in real-time. The algorithm detects the faulty device correctly for a large system, and its performance shows 97% accuracy, with a short computation time (Natarajan et al., 2020). However, this technique is PV sensor-specific and is unsuitable for solar sensors.

Another work discovered anomalies in the base station's recorded fuel consumption data. Abnormalities were found by learning fuel consumption patterns using four classification methods: logistic regression (LR), SVM, KNN, and multilayer perceptron (MLP), performing the normal feature engineering, selection, fitting the model classifiers, and then testing. According to the findings, MLP and SVM were the most efficient and yielded a score of 96% and 95%, respectively (Mulongo et al., 2020). However, the work used only one data source and did not consider streaming data or handling missing values.

Unsupervised Algorithms

Similarly, various unsupervised solar sensors' malfunction detection algorithms were introduced in the literature, achieving good performance. Mainly, they are based on artificial neural networks (ANN) and other variations, e.g., recurrent neural networks (RNN) and convolutional neural networks (CNN). The advantage of using these techniques is that they work well for process automation and handle time series data well, making them beneficial for the problem of this research work.

(Sanz-Bobi et al., 2012) This paper presented a new tool (ISDIPV) capable of continuously and automatically detecting and diagnosing anomalies in solar devices and reporting them immediately. The two modeling methods described the expected performance: linear transfer functions (LTF) and NN models based on MLPs. This is suitable for solar sensors that are distributed and operated

unattended, like this intended research case. However, it works only for small and mid-sized systems and has weak interpretability for the variation of the solar radiation measurements.

An essential type of anomaly is contextual, representing some solar data anomalies such as specific measurements that can be acceptable during dusty weather but not during the clear sky. Kosek, 2016 proposed a new ensemble contextual AD approach for cyber-physical intrusion detection in smart grids using an ANN to recognize the behavior of a distributed energy resource. The model outperforms point AD by 55% in control detection and 56% in malicious control detection on average (Anna Magdalena Kosek, 2016).

Although this work was based on one data source, voltage control, it gives a foundation for NNs to identify anomalies; hence, DL will perform better.

In the same context, different types of anomalies common in solar data are introduced by (Rossi et al., 2016), i.e., the collective anomaly, which refers to groups of occurrences that may be abnormal based on their appearance patterns. A large AD study was conducted using unsupervised contextual and collective data flow by a large energy distributor. The method investigated distinct potential abnormalities. The categorical clustering and frequent item-set mining techniques were combined for accurate AD. Then, a visual depiction of the possible anomalies and a list of the top ten anomalies were presented. However, the proposed work used datasets from a different domain, required a domain expert's opinion, considered itemset instead of sequences, and did not consider streaming data and real-time system behavior.

Another study implemented an abnormality prediction and alert mechanism to indicate the need for predictive maintenance for the possibly affected solar device. The model was built based on an ANN to predict AC power generation using solar irradiance and the temperature of the measurement sensor. The approach offers an AD positive rate above 90% (De Benedetti et al., 2018). This can help as the data correctness can be affected by an issue related to the sensor, and it will effectively alert operatives to act immediately. However, the work used a comparison approach between the measured and predicted values that did not consider factors like seasonality and weather conditions, which can lead to inaccurate AD in solar data.

The paper (Pereira & Silveira, 2018) presented a general, unsupervised, and scalable scheme for AD in time-series data that could be run offline and online. The scheme was built around a rebuilding model that used a variational AE. The encoder and decoder were parametrized with RNN to recognize the temporal reliance of the time-series data. The model could detect anomalies employing probabilistic restoration metrics such as anomaly scores. Also, it offers the understanding of detecting data dependencies and handling real-life online data, which is pertinent to the continuously collected and stored solar data. However, the work only addressed the univariate time series data, and the normality definition is relative and may not apply to a different context.

The study by (Hempelmann et al., 2020) focuses on AD with Image noise reduction and segmentation post-processing procedures. Distinct models were selected to examine the energy yield data. The results show that the best detection rates were achieved using NN, linear, probabilistic, proximity-based models, and anomaly ensembles. However, the model detects only new occurring faults but cannot detect existing long-term defects like shadows.

This work proposed a technique for AD automation considering domain-specific requirements and corresponding constraints, such as online monitoring, offline certification, and live monitoring of data acquisition rates. The authors introduced a classifier to detect known abnormal behaviors and expand current monitoring coverage to detect new failure modes for sensors online monitoring for early alert. The CNNs algorithm was efficient on currently tracked failure modes. Furthermore, using offline monitoring, AEs identified emerging problem behaviors on numerous low-resolution observations. Also, it highlighted the cruciality of short-term solar energy models for AI-driven Internet of Things (IoT) modeling (Pol, 2020). However, the work did not consider external feature variables, e.g., seasonality, presented difficulties in the traditional forecasting method, and classified all the data without any human intervention, while this intervention is required in some cases.

This research used DL to investigate an image-based method for automating the manual AD process on quality control plots. They used images of time series quality plots to train a proposed CNN model to accomplish this. The 1-class output model learned to identify plots with anomalies like those detected by a human reviewer with a high classification score. Using Gradient-based Class Activation Mapping, they could successfully identify and color-highlight the classified anomalous regions within the quality plots. Accordingly, they obtained a 0.92% overall F1 score for anomaly classification and 91% accuracy for anomaly localization (Srinivasan et al., 2020). However, the work used weekly data resolution and avoided manual efforts, leading to low FP tolerance and the inability to classify other error types.

(Zhou et al., 2021) proposed a hybrid DL method that combined clustering techniques, CNN, long short-term memory (LSTM), and attention mechanism (AM) with a specific sensor type. The proposed method was divided into three stages: clustering, training, and forecasting. Correlation analysis and self-organizing mapping were used at the clustering stage to select the most relevant factors from historical data. Then, to perform the forecasting task, a CNN, LSTM NN, and an AM were combined in the training stage to create a hybrid DL model. Next, the appropriate training model was chosen during the testing stage based on the month of the testing data. The method's prediction accuracy rates were high for all time intervals but not higher than other individual models.

According to (M. Kardi, 2021), A DL-based approach for detecting anomalies in electricity consumption data was presented in this article. This problem was addressed in two stages: (a) create a neural network model adopting LSTM for sample prediction, and (b) use the output as input for another LSTM autoencoder to learn the characteristics of normal consumption. The study considered several weather features, including pressure, cloud cover, humidity, temperature, wind direction and speed, and temporal and lag features. Therefore, the local and global anomalies were distinguished, and a feature selection method was implemented to find the optimal features for good results. The results revealed that the temporal and lag features improved the proposed method's efficiency and performance. However, it was difficult to assess the performance of AD. Also, there was no labeled data, and the work used hourly data resolution.

(Faber et al., 2021) introduced a fully automated multivariate time series AD framework for failure prevention, reducing repair costs and losses. This model is the first to apply an ensemble DL AD model using a neuroevolution strategy to boost the AD scores of different models. The results showed that the framework could boost most AD models based on DL in a reasonable time. Yet, the model used inadequate sensors and samples for training and testing.

(Sharif et al., 2022) proposed a framework scalable in time series data for AD based on an unsupervised technique. The proposed approach is based on a variational AE, a deep, productive model incorporating variational belief with DL. Also, real-time data analysis was performed using LSTM networks to process, predict, and classify based on big time-series data. The findings prove that all parameters were trainable, and a dropout layer bypasses the overfitting issue. Nonetheless, it was difficult to assess the performance of AD.

Ibrahim et al., 2022 could assess the performance of various ML schemes and use them to detect anomalies in solar sensors. AE-LSTM, Facebook-Prophet, and Isolation Forest were among the schemes tested. The AE-LSTM was the most accurate model in distinguishing normal and abnormal PV system behavior. The correlation coefficients between the plant's internal and external feature parameters were calculated and used to assess the effectiveness of ML models in detecting anomalies (Ibrahim et al., 2022). However, the model does not apply to large-scale systems.

Another study proposes a framework for automatic sensor failure detection and diagnosis (FDD) with notification capability to maintain the quality of monitoring systems. The method uses unsupervised algorithms to detect variances in energy power production to provide three innovative models:

- Time series forecasting utilizes the prophet procedure that performs best with substantial seasonal effects and data from several seasons.

- ANN with month character.
- Hour vote.

The model discriminates abnormalities and yields a balanced ACC, an acceptable AD rate of 0.736, and an MCC rate of 0.863, respectively. Further, the result supported the previous conclusions that the classification score of the MLP regressor outperformed the other ML models (Klinsuwan et al., 2023). This approach decreases complexity and examines each day individually to discover abnormal behaviors, but it employs a single model for all data.

Semi-Supervised Algorithms

The semi-supervised algorithms were used in literature to predict the failure of the solar sensors. (Tsai et al., 2020) proposed an AD technique based on a semi-supervised learning model to predict equipment conditions to assure continuous operation and precise solar panel power production. This method built the classifier using the AE neuron network model and the clustering technique for normal incident filtration. However, the technique used only small and medium-sized solar station datasets and cannot detect insensible anomalies as the dataset sampling rate is too low.

(Oprea et al., 2021) analyzed large datasets collected from smart meters installed in a trial study in Ireland by applying a hybrid approach. The proposed framework detects abnormal values in the time series using an unsupervised ML technique, establishes a threshold for the percentage of abnormal values from the dataset, and labels the data as normal or suspicious. The work proposed two algorithms for AD for unlabeled data: SR-CNN and an anomaly-trained model based on martingales to determine variations in data streams. Next, the researchers applied the Two-Class Boosted DT and FL Discriminant analysis to the previously processed dataset. The approach achieved a 90% accuracy in detecting suspicious values, a precision score of 0.875, and an F1 score of 0.894. However, the method addresses only the univariate data and needs to be validated by testing more complex and diverse datasets.

Data/Model-Driven Methods

A different conceptual method for AD is data/model-driven, based on data from similar nearby sensors, historical data in the same place, or several sensors in the exact location. Data-driven techniques became popular for complex system detection, diagnostics, and prognostics.

A data-driven method, SolarClique, detects anomalies in a solar establishment's power generation without any specific requirements related to the device type or weather-related supplemental data. Instead, the method utilizes the data of nearby sites to detect anomalies in the target site data. The method validation shows high accuracy, even with few nearby stations; hence, it can be scalable. Also, the method can identify the source of data variations from anomalies or other factors like weather conditions (Iyengar et al., 2018). However, the method did not consider the seasonal component and used hourly data resolution.

The study of (Zhao et al., 2019) employs another method with the same data-driven concept for AD. The AD and classification use sensors' string currents as indicators to reveal and classify solar system anomalies. In addition, the proposed work consists of two methods: hierarchical context-aware AD (unsupervised ML technique) and a multimodal anomaly classification method. The validation process showed robustness, effectiveness, and cost and time efficiency (Zhao et al., 2019). However, there were data limitations (availability and resolution); the method considered only five weather conditions, excluding dust, and did not consider other factors (device misalignment, cleaning).

(Correa-Jullian et al., 2019) investigated similar concepts using DL algorithms such as ANN, RNN, and LSTM in a solar hot water (SHW) system. Nevertheless, continuous monitoring of the systems is required to avoid unrecognized systems malfunction, which will increase the system's total cost. The results of temperature predictions using the different algorithms were similar. However, the LSTM models achieved a more reliable performance with the lowest values of 1.27°C for RMSE,

0.55°C for MAE, 0.52 °C2 for variance, and 3.45% for relative prediction error. Additionally, under various meteorological conditions, the AD model achieved a mean accuracy of 85% and 82%, respectively (Correa-Jullian et al., 2019). However, the method used synthetic data while historical data is needed for accuracy, isolated and study limited behaviors and anomalies, and no anomalies classification was implemented.

A model-driven method, SunDown is designed for detecting per-panel faults in solar devices without installing additional new sensors. Instead, the model discovers deviations from the anticipated performance by correlating the output of the devices in the array. Also, the model manages concurrent failures in multiple panels, identifying anomalies and determining their sources. The model showed a mean absolute percentage error (MAPE) of 2.98% for prediction (Feng et al., 2020). However, the work did not use solar faults datasets, considered only three faults' sources, and leveraged correlations between devices in the array. Yet some solar stations have inadequate devices at their sites. Nevertheless, since the ground solar stations in Saudi Arabia, the main target of the study, have some redundant sensors and other stations nearby in most sites, these works can inspire our framework.

Tailored Algorithms

(Branco et al., 2020) proposed five tailored algorithms to detect several sensor anomalies, such as shading and orientation. The detection algorithms processed several time series samples obtained for two periods with different weather conditions. The results suggested that most time-series labels were successfully detected under favorable weather conditions with a low percentage of false positives. The proposed method can benefit energy utilities and owners of solar systems in monitoring the operation and performance of their systems easily. The study limitations were using datasets from small solar stations, algorithms designed for PV sensor-specified faults, inadequate weather conditions, and low prediction accuracy under adverse weather conditions.

This study is different because the algorithm was developed manually to suit each case. Thus, it offers a different perspective, as dealing with big solar data affected by other conditions requires a tailored technique. Further, finding a single straightforward technique to address all anomaly types is not viable.

Electroluminous (EL) Technique

Another different conception is the Electroluminous (EL), an interesting technique discussed comprehensively in many studies to reveal the effect of potential sensor defects. This is related to this review as the solar data consists of some images (i.e., camera devices and data quality graphs). (Fuyuki et al., 2005) introduced in this work the EL technique, which can detect various sensor issues that other methods would miss. It is based on an optical phenomenon in which a photovoltaic cell material emits light at a specific wavelength peak in response to an electric current that stimulates and reverses the operation of the sensor. According to (Mochizuki et al., 2016), the light is captured by a special camera outfitted with special optical filters to obtain the electroluminescence image of the PV cell. The EL procedure was considered one of the least invasive and expensive methods. In addition, the resulting images were of high quality, allowing for detecting subtle anomalies. However, capturing EL images required specific conditions to avoid residual radiation emission. Thus, the availability of correctly annotated public EL datasets was limited (Mochizuki et al., 2016).

Using data extracted from high-resolution EL intensity images, two approaches were investigated for automatically detecting device defects. The approaches classify data using a Support Vector Machine (SVM) and an end-to-end deep CNN. The result shows that CNN outperforms the SVM with an average accuracy of 88.42% compared to 82.44%. Both automated methods provide continuous, accurate monitoring of PV devices (Deitsch et al., 2019). However, CNN requires high hardware specifications.

EL is the only method for detecting non-electrical active cracks at the cell level, making it the preferred approach for AD tasks. Also, the characteristics of the EL method make it usable to identify

other devices' defects. The paper proposed a method for combining the electrothermography (ET) and EL detection effects of defects. This research looked into electromagnetic induction (EMI) and image fusion, showing that EMI significantly improves ET and EL's detection capability by combining their two wavelengths (Yang et al., 2020). Moreover, the additional capabilities of the combination of EL and ET confirm the initial assumption that no single method can work efficiently alone. However, this method had no AI algorithm to help with automatic defect classification (for fast detection, image display, and online monitoring) and required specific hardware.

Furthermore, the work proposed in (Mayr et al., 2020) used EL approaches to segment anomalies in electroluminescence imagery using unsupervised learning techniques such as semi-supervised or weakly-supervised learning strategies. This method aggregates the activation maps into single scores for classification. The technique accurately identified cracked solar sensors using small training examples and image-level annotations, which might contribute to the effective production and maintenance of solar projects and develop novel weakly supervised segmentation algorithms. However, it is hard to decide the best segmentation (only a rough cell segmentation from image-level labels).

More attempts have been made to detect anomalies using EL imagery (Tang et al., 2020). However, most of these approaches relied on undisclosed private data or used the dataset (ELPV). However, the EL procedure is preferred for detecting structural defects, intrinsic defects (such as fingerprint marks), and extrinsic defects (such as micro-cracks, cell degradation, corrosion, and electrically isolated parts), according to the manufacturer (Rahman et al., 2021).

A framework for detecting abnormal compositions in solar sensors using EL images was proposed by (Julen Balzategui, 2021). The AD approach was robust, enabling the detection of variances in two stages: (1) uses a generative adversarial network (GAN) to train a small number of non-defective samples; this is unique as GAN is known for being difficult to train, (2) the automated generated features, utilized to train a supervised model, CNN, for discovering various failures. The segmentation and classification results achieved by supervised models trained with automated labels are equivalent to those attained from manual labels. This study has value for the quality inspection process as it reduces the network's training instability, resulting from the cohesive components included in most quality inspection instances. However, the inspection cost saving is traded with hardware requirements cost. Additionally, the model requires some hardware specifications for unsupervised training, and the diverse architectures and parameters must first be investigated for method optimization.

Another study proposed an automated pipeline for EL image processing. The images were classified as solder disconnections, oxygen-induced, cracks, and intra-cell defects. The methods are posted as open-source software. The results show an excellent detection of the four types of anomalies, and the method can be used and modified for other types of anomalies (Chen et al., 2022). However, various observed EL images for each defect type are needed, EL images from other sensors are limited for generalization, and annotations showed low quality.

(Rahimzadeh et al., 2022) applied an advanced DL method, PatchCore, a visualization-based AD, to identify defective sensors from their EL images. The framework considers giving each image to the same deep NN in the training and testing phases and comparing the extracted and typical features to identify the deviations. The implementation showed a high performance in detecting anomalies with an AUROC score of 0.97, with a smaller training dataset. Therefore, it is versatile and efficient for the quality control of new sensors' production. However, this method requires cost and knowledge to automate the defect detection processes for EL images efficiently.

DL techniques for automatic AD in Electroluminescence (EL) images were also the focus of Otamendi's (2022) research. Automated anomaly annotations assist in constructing decision-making systems (DMSs) to detect failures, prolong the sensor life, and enhance the operation and maintenance processes. The authors proposed a benchmark using a mix of advanced data-driven methodologies for anomaly segmentation annotations. The method demonstrated flexibility to new sensors, cost-effective refinement, and advanced annotations generated using public datasets. Also, the approach was validated by annotating a commonly used dataset, resulting in a 60% decrease in annotation

cost (Otamendi et al., 2022). However, it requires a tuning process once per dataset. Additionally, it needs high-level annotations to enhance the ability to train supervised models, increase accuracy and performance, and reduce manual revision time.

Forecasting Methods

Forecasting is a key function in RE. Forecasting and AD models are similar since both detect and label unknown data. Further, both techniques use DM and machine learning algorithms and work with big data.

(Liu, 2018) investigated forecasting models such as the LR, the Lasso regression, the Ridge regression, the SVR, and the MLP models. The forecasting model was tested and evaluated using EVS, MSE, R2, and processing time and is suitable for real-time monitoring systems. The Linear, Lasso, and Ridge regression models are advantageous for the short time training and testing data on the cost of lower prediction accuracy. Conversely, the MLP model shows the best training and testing data time and accuracy. However, the weather data for prediction is from a different station far from the targeted station.

(Malakar et al., 2021) discussed the short-term forecasting of GHI because of its significant impact on incident solar radiation. The work used data from three solar stations in two climatic zones over two seasons. The model's forecasting accuracy outperformed three recent benchmark approaches based on RF, RNN, and LSTM. In addition, they confirmed the significance of considering the temporal data order, the lack of any benefit from data preprocessing, and the impact of making the LSTM model stateful. Also, the variability of the input data influenced the complexity of an LSTM network and batch size. However, the prediction is only for the short-term, and the implementation is limited to two climatic zones.

This work built 12 ML models for predicting daily and monthly solar radiation values. Then, a stacking model based on the best of these algorithms (GPR, GBRT, XGBoost, and RF) was developed for prediction. The stacking model outperformed the single models in daily prediction but had no benefit over the XGBoost model in the monthly prediction. Also, the climatic parameters (such as temperature, visibility, and sunshine duration) are critical in ML models. It was concluded that the stacking and XGBoost models best predict solar radiation (Huang et al., 2021). The work illustrates the idea of obtaining a higher performance by assimilating advanced learners using multiple lower-level learners. Therefore, as it will be difficult with such big data to train the entire data at once, it might need to split the data into small sets. However, models using daily and monthly resolution had a low performance for the monthly model due to scarcity of training data and showed low performance with large data sizes.

(Khan et al., 2022) created a lightweight ESNCNN RE prediction model where an echo state network (ESN) learned the nonlinear mapping relationship and a CNN extracted spatial information from RE data. The combination of ESN and CNN improved ESNCNN characteristics for prediction. The model's performance is compared to various advanced approaches. The ESNCNN's generalization ability reveals a significant decrease in error rates compared to other advanced approaches (MSE (5.01%), MAE (5.49%), and RMSE (3.76%)). However, the model used a five-minute resolution, showed low detection accuracy, and did not consider different sources of images and additional weather parameters to develop a more effective prediction model.

Table 1 offers a summary of the literature reviews.

DISCUSSION

In this research, we consult the literature to explore all available techniques to select and design a single system that can solve the research problem of how to confirm data correctness. We focused on the implemented comprehensive methodologies for accurately predicting solar sensors' failures, which overcomes the challenges of rapidly obtaining massive time series data. It is essential to validate

Table 1. A summary of the literature reviews

References	Taxonomy	Technique(s)	Contribution	Performance	Limitation(s)
Toshniwal et al., 2020	General surveys	NN, SVM, BN, Rule-based, KNN, Relative Density, Density-Based Clusters, Statistical Distribution, spectral, Information Theoretic, graphing methods	The best AD method for a specific problem depends on criteria (input data, type of anomalies, output data, and domain expertise).	-	Dataset for AD must consist of the majority of nominal instances to identify rare events.
Hu, 2012		ARIMA, SVM, NN, k-NN, CNN, LR	Successfully detect ten types of anomaly sources, while all previous algorithms detect maximum five types.	K-NN classification manages to label correctly 97% of testing anomalies.	Algorithms perform well with small test cases, address only univariate time series data.
Himeur et al., 2021		1-class and traditional classifications, dimensionality reduction NN, regression, probabilistic models, feature extraction strategies (distance, density, graph-based, time-series analysis, compressive sensing)	No method discovers all types of anomalies using low-dimensional subspaces.	-	Scalability, easy implementation, and privacy preservation are not considered, unsupervised learning Models can only detect one type of anomaly, supervised methods require labeled datasets to learn the abnormalities.
Moritz Benninger, 2019	Supervised Algorithms	KNN and a one-class SVM (OCSVM)	The learning algorithms considerably reduce the measuring effort and simultaneously offer reliable anomaly monitoring	-	Hardware requirement (contactless sensors)
Harrou et al., 2019		1-class support vector machine (1-SVM)	The technique has high detection efficiency with actual data, quantifies dissimilarity between normal and abnormal features, handle nonlinear features, and makes no assumptions on data distribution.	Technique outperforms unsupervised conventional binary clustering algorithms (i.e., K-means, mean-shift, EM, Birch, and agglomerative)	Address only anomalies of temporary shading, built based on the one-diode model (PV sensor specific) and unsuitable for the solar sensors.
Natarajan et al., 2020		Thermal image processing, SVM tool	Detects the faulty device correctly for a large system, Extracts features in real-time with significantly less time complexity.	Performance shows 97% accuracy, with a short computation time.	Validation used a single factor only.
Mulongo et al., 2020		LR, SVM, KNN, MLP	MLP and SVM were the most efficient in the measurement assessment based on the data type and the underlying assumptions employed.	MLP and SVM yielded an accuracy score of 96% and 95% and recorded AUC of 0.98 and 0.96 for a precision-recall curve	Did not consider streaming (online) data, does not handle missing values, only one source of data from the main station.
M.A. Sanz-Bobi, 2012	Unsupervised Algorithms	linear transfer functions (LTF), NN models based on MLPs.	Continuous and automatic detection and diagnosing of anomalies in solar sensors and reporting directly, redundancy between models ensures reliable and robust AD.	-	Only work for small and mid-sized systems, weak interpretability for the variation of the solar radiations.
AM Kosek, 2016		ANN	Satisfying AD that outperforms a single point AD.	Outperforms point AD by ~55% on average.	Based on voltage control, only gives a base on NNs AD.
Rossi et al., 2016		Combined categorical clustering and most frequent itemset (MFI) mining techniques	Consider contextual and collective data flow by a large energy system, a visual anomalies depiction for stakeholders, a list of the top ten anomalies.	-	Datasets used are from a different domain, requires domain expert's opinion, considered itemsets instead of sequences, did not consider streaming data and real-time system behavior.
De Benedetti et al., 2018		ANN	Successfully recognize data deviations and issue alarms for predictive maintenance, helping validate data correctness by revealing sensor issues.	AD positive rate above 90%.	Used a comparison approach between the measured and predicted values, which is different for solar data as the seasonality affects values and weather cause data fluctuation.

continued on following page

Table 1. Continued

References	Taxonomy	Technique(s)	Contribution	Performance	Limitation(s)
Pereira & Silveira, 2018	Unsupervised Algorithms	Variational Bi-LSTM autoencoder, RNN	Detect data dependencies and anomalous on real-life online data by employing probabilistic restoration anomaly scores.	-	Only address the univariate time series data, the normality definition is relative and may not apply to a different context.
Hempelmann et al., 2020		LSCP, VAE, or MOGAAL	Addresses faults with little data or may occur in the future, detect the actual behavior of the sensor.	Highest performance rate ~92% for VAE detecting faults, ~92% for MOGAAL detecting Shade.	Only detect new occurring faults and not existing long-term defects like shadows.
Pol, 2020		CNN, AE	Algorithms based on AE offer a robust AD strategy, used for a semi-supervised novelty detection on several low-resolution samples of offline monitoring.	CNNs are highly efficient in improving speed of flagging emerging faults and scalability compared to classical models.	Do not consider external feature variables such as seasonality, classify all the data without any human intervention, while it is required in some cases.
Srinivasan et al., 2020		A CNN, DL, Visual Geometry Group	Identify time-series quality plots that contained anomalies similar to those a manual reviewer would detect with a high AD score.	Overall recall score for AD of 0.95, F1 score of 0.92, the accuracy of 91%, and precision of 0.88.	Use weekly data resolution, FP tolerance is too low due to avoiding manual efforts, 1-class output model (cannot classify different error types for other variables).
Zhou et al., 2021		Clustering techniques, CNN, LSTM, attention mechanism (AM), hybrid DL	Higher prediction accuracy rates than traditional ANN, LSTM neural network, and a combination of LSTM NNs and AM	High prediction accuracy rates for all time intervals of 7.5 min.	The accuracy achieved is not higher than other individual models.
M. Kardi, 2021		AE-LSTM, feature selection, RNNs, DL	Considered several weather features, including pressure, cloud cover, humidity, temperature, wind direction and speed, and temporal and lag characteristics.	The temporal and lag features improved the method's efficiency and performance.	Difficult to assess the performance of AD, no data is labeled, use hourly data resolution.
Faber et al., 2021		Ensemble DL, three models of the AE, CNN, VAE contain two LSTM layers, utilizes the idea of AE and GANs architecture.	The first approach built an ensemble DL AD model fully automated using a neuroevolution strategy, efficient AD process in a reasonable time.	Good performance by all models, except graph NN models, CNN 1D gained the best results.	No enough number of sensors and samples for training and testing.
Sharif et al., 2022		DL, VAE, AE-LSTM, LSTM-NNs	Successful AD methods combine variational belief with DL real-time big-time-series data analysis using LSTM.	All parameters were trainable, and a dropout layer bypasses the overfitting issue.	Difficult to assess the performance of AD.
Ibrahim et al., 2022		AE-LSTM, Prophet, and Isolation Forest (IF)	AE-LSTM was the most accurate in distinguishing normal and abnormal PV system behavior.	AE-LSTM correctly identified anomalies, Prophet has 53 FP, IF mislabeled all signal peaks.	Low prediction accuracy for Prophet and IF because of their sensitivity to noisy signals, does not apply to large-scale systems.
(Klinsuwan et al., 2023		Prophet, ANN, hour vote, MLP regressor	MLP regressor gives the best classification score, outperforming the other machine learning models, decrease complexity.	The model yields AD rates of 0.736 and 0.863 for MCC and a balanced ACC, respectively.	Employs a single model for all data from several seasons.
Tsai et al., 2020	Semi-supervised Algorithms	AE, NN model, clustering technique	AD method uses a semi-supervision learning model and clustering technique to filter normal incidents to detect faults.	The mechanism has a certain extent of feasibility.	The dataset from small and medium-sized solar stations, can not detect insensible anomalies as dataset sampling rate is too low.
Oprea et al., 2021		Spectral Residual-CNN (SR-CNN), an anomaly-trained model based on martingales, 2-Class Boosted Decision Tree, Fisher Linear Discriminant analysis	Successful hybrid architecture of different ML models for real-time predictive analytics; unsupervised detect unlabeled data, supervised models for classification.	Detect anomalies with an accuracy of 90%, a precision score of 0.875, and an F1 score of 0.894.	Requires testing on more complex and various datasets, only address the univariate data.

continued on following page

Table 1. Continued

References	Taxonomy	Technique(s)	Contribution	Performance	Limitation(s)
Iyengar et al., 2018	Data/Model-Driven	KNN, SVR, DT, RF, LR	Robust and scalable methodology and can identify the source of abnormal events by utilizing the data of nearby sites.	Achieved high accuracy, even with few nearby sites.	Use hourly data resolution, did not consider the seasonal component.
Zhao et al., 2019		Auto Gaussian Mixture Model GMM clustering, XGB, SVM, BGG classification	Multimodal (Hierarchical context-aware) unsupervised AD method, accurately detects and classifies various solar system anomalies.	Model is robust and effective with efficient cost and time, the accuracy of XGB 97%, BGG 95%, and SVM 80%.	Data limitation (availability and resolution), Considered only five weather conditions excluding dust, did not consider other factors (device misalignment, cleaning).
Correa-Jullian et al., 2019		DL algorithms, ANN, RNN, LSTM	Successful predictions under various climatic conditions using the different algorithms, but LSTM models achieved the most reliable performance, while DNN model has the lowest performance.	AD mean accuracy of ~84%, LSTM had 1.27°C for RMSE, 0.55°C for MAE, 0.52 °C2 for variance, 3.45% for relative prediction error.	Used synthetic data while needing historical data for accuracy, isolated and study specific behaviors and anomalies, no anomalies classification implemented.
Feng et al., 2020		LR-Based Model, Graphical Model, Half-Sibling Regression, RF classifier	Two model-driven techniques successfully manage concurrent failures in multiple panels, identify anomalies, and determine sources.	The model shows MAPE of 2.98% for prediction per device, and accuracy >97% classifying faults.	No datasets of solar faults are available, consider only three faults' sources, leverages correlations between devices in the array, but solar stations have mostly two devices at a site
Branco et al., 2020	Tailored Algorithms	five algorithms are based on simple rules	Detect most time-series labels under favorable weather conditions with a low percentage of FP, help in monitoring the systems' operation and performance.	FP $\leq$ 30 for the different models showing low reliability and cost for solar projects.	The dataset from small solar stations, algorithms designed for PV sensor-specified faults, did not consider diverse weather conditions, low prediction accuracy under adverse weather conditions.
Takashi Fuyuki, 2005	EL	-	EL techniques can detect various sensor issues that other methods would miss.	-	-
Mochizuki et al., 2016		-	Consider one of the least invasive and expensive methods, allow detecting subtle anomalies.	The resulting images were of high quality.	The availability of correctly annotated public EL datasets is limited, as accurately capturing EL images requires specific conditions.
Deitsch et al., 2019		SVM, an end-to-end deep CNN	Automated methods provide continuous, accurate sensors monitoring, SVM can run on arbitrary hardware.	CNN outperforms the SVM with an average accuracy of 88.42% to 82.44%.	High hardware requirements by CNN.
Yang et al., 2020		Combining electrothermography (ET) and EL complementing the two wavelength detection data, electromagnetic induction (EMI) and image fusion rule	AD model and image fusion system integrated ET and EL based on EMI, helps identify other sensor defects, different modality images by diverse sensors.	No single method can work efficiently alone, the algorithm is superior to others.	Hardware requirements, No AI algorithm to help automatic defect classification (for fast detection, image display, and online monitoring).
Mayr et al., 2020		a weakly supervised segmentation approach using DL	Contribute to the effective production and maintenance of devices and development of novel weakly supervised segmentation algorithms, segmentation performance is not strictly related to that of classification.	Accurately identifies faulty solar sensors with few training samples and image-level annotations, with one forward pass through the network.	A rough cell segmentation from image-level labels only, hard to decide the best segmentation.

continued on following page

Table 1. Continued

References	Taxonomy	Technique(s)	Contribution	Performance	Limitation(s)
Tang et al., 2020	EL	DL using GAN, CNN	Combines traditional image processing technology and GAN characteristics, generating a large number of high-quality EL images for limited EL image samples.	An efficient model for automatic defect classification with the generated EL image.	EL image generation method does not apply to all sensors' types or defects, performance varies depending on the quality and resolution of EL images and complexity and diversity of the defects, evaluation is limited to a few real-world scenarios, technical expertise and hardware requirements.
Rahman et al., 2021		Main factors that degrade solar systems, best techniques for characterizing defects, and their advantages and limitations.	EL procedure is preferred for detecting structural, intrinsic, and extrinsic defects.	-	-
Julen Balzategui, 2021		GAN, CNN	f-AnoGAN-256 is the best model, valuable for the quality inspection process as it reduces the network's training instability, resulting from the cohesive components in most quality inspection cases.	AD has a high accuracy rate with reduced processing time, segmentation and classification trained with automated labels equivalent to manual labeling.	Needs optimization for architectures and parameters, hardware requirements (require 2 GPUs for unsupervised training), inspection cost saving is traded with hardware requirements cost.
Chen et al., 2022			Excellent detection of different anomalies types, and can be used and modified for other classes, methods made as open-source software to analyze EL images.	Models perform similarly well, with macro F1 scores on a real-world testing set of 0.83 (ResNet18) and 0.78 (YOLO).	Consider only four types of anomalies, need a variety of EL images observed for each defect type, limited EL images from other sensors prevent generalization, low annotations quality
Rahimzadeh et al., 2022		DL	Visualization-based deep NN that is versatile and efficient for the quality control of new sensor's production.	High performance in AD with an AUROC score of 0.97, with small training dataset.	Cost and knowledge requirements to automate the AD processes for EL images efficiently.
Otamendi et al., 2022		Weakly-supervised DL model, a segmentation annotations technique, unsupervised clustering (DBSCAN)	Combines data-driven models to extract anomaly segments' annotations from EL images across various solar sensors' types, cost-effective and efficient with small and general datasets, uses DL methods to build the process, and creates DMS for AD.	Validation by annotating a commonly used dataset, resulting in a 60% decrease in annotation cost.	Requires a tuning process once per dataset, needs high-level annotations to enhance the ability to train supervised models to increase the accuracy and performance and reduce manual revision time.
Liu, 2018	Forecasting Methods	LR, Lasso regression, the Ridge regression model, SVR, MLP	Five models are used in real-time monitoring systems. The Linear, Lasso, and Ridge regression models are better for short processing time training/ testing data, on the cost of lower prediction accuracy.	MLP model is the best performance considering the training and testing data processing time and accuracy.	Low detection accuracy, weather data used for predicting is from different and far from the targeted station.
Malakar et al., 2021		DL, LSTM, DSS-LSTM, DSSL-LSTM	Predicting of GHI using DL sequence model with stateful/ stateless LSTM. Confirms significance of considering temporal data order, reveals impact of the stateful LSTM (DSS-LSTM) model, found that input data variability influenced total nodes in LSTM network and batch size.	DSS-LSTM outperformed three recent benchmark methods based on RF, RNN, and LSTM with a substantially lower value of nRMSE 2.25%.	Prediction is only for the short-term, the implementation is limited to 2 climatic zones.

continued on following page

Table 1. Continued

References	Taxonomy	Technique(s)	Contribution	Performance	Limitation(s)
Huang et al., 2021	Forecasting Methods	GPR, GBRT, XGBoost, RF, ELM, DT	12 ML models to predict daily/monthly solar radiation values and stacking model based on the best of these algorithms (GPR, GBRT, XGBoost, RF). Stacking model outperformed single models in predicting daily solar radiation but is similar to XGBoost model in monthly prediction.	The stacking and XGBoost models best predict solar radiation (XGBoost ~ 0.922 of R2), while ELM and DT models had the poorest precision. RF had a higher time complexity and XGBoost had the shortest.	Use daily and monthly resolution, low performance for monthly model due to very little training data, low performance with large data size.
Khan et al., 2022		ESN, CNN layers are connected linearly via residual connections.	ESNCNN learns the spatiotemporal features of the refined input data more effectively than other models, with a lower time complexity for testing and training.	Generalization of ESNCNN reveals a significant error rate reduction compared to other approaches, MSE ~5.01%, MAE ~5.49%, RMSE ~3.76%.	Use five minutes resolution, low detection accuracy, did not consider different sources of images and additional weather parameters to develop a more effective prediction model.

collected solar data that reflects the device's healthy performance. This is because the energy systems and acquired data are vital for many crucial energy projects, including forecasting solar system capabilities in measuring solar radiation.

This paper discussed various studies examining different AD techniques in the context of big data and energy systems, focusing on solar sensors. These studies were categorized under eight main groups concerning the content presented, including general AD surveys, AD with supervised, unsupervised, and semi-supervised, AD based on data or model-driven concepts, Tailored algorithms for AD, AD in images using EL concept, and solar energy forecasting as another face for AD.

The main implications and takeaways from the literature are:

- The choice of AD techniques for big energy data depends on several factors, including the nature of the data and the specific problem description. There is no one-size-fits-all solution.
- Supervised learning algorithms can perform well in AD, but SVM is noted for its memory efficiency with high-dimensional data. However, their performance may vary for time series big data like solar energy data.
- Unsupervised approaches offer successful sensor failure detection results. These algorithms address various aspects of sensor failures, including offline and online AD, data dependency detection, and real-life situations. They are useful for anomaly detection in continuous data collection scenarios.
- Semi-supervised techniques that combine supervised and unsupervised algorithms, such as neural networks, autoencoders, boosted decision trees, and Fisher linear discriminant, can be beneficial when labeled data for training is limited. Feature extraction algorithms and ensemble methods can enhance performance.
- Some approaches focus on comparing data from the same device or similar devices rather than relying on external sources like weather data. These concepts can be effective but are most accurate when dealing with a limited number of sensor failure sources.
- EL is valuable for detecting various device failures, including structural, intrinsic, and extrinsic defects, especially when dealing with image and graph data. Combining supervised, unsupervised, and semi-supervised algorithms in EL can yield superior results.
- AD models share similarities with forecasting models, especially in real-time monitoring systems. Regression-based models are efficient in processing time, while MLP-based models

offer high accuracy. Temporal data order is essential, and data preprocessing may not always provide significant benefits.

- DL techniques offer improved accuracy, flexibility in detecting new sensor types, and automation capabilities for AD in energy data. They are particularly effective when combined with other approaches and can be fine-tuned with cost-effective annotations using public datasets.
- A notable challenge is the scarcity of segmentation annotations for solar systems. This limitation hinders the development of anomaly segmentation approaches, leading to using semi-supervised and weakly supervised techniques. More research is needed in this area to address systematic and weather-related variations.
- Building AD models for solar data should consider other factors: location, elevation, and weather conditions. However, existing literature tends to focus on specific issues individually, and there is a need for more comprehensive approaches that consider multiple factors simultaneously.

Although the review provides a comprehensive overview of the research in this area, it is essential to acknowledge some limitations:

- The literature mainly tackled only one issue at a time, and some issues were not covered, such as data variations due to instrument position (leveling) and cleaning, among other factors.
- Many of the discussed papers report the performance of their AD techniques, but they often lack comprehensive evaluation metrics. Metrics such as precision, recall, F1-score, ROC curves, and AUC are crucial for understanding the trade-offs between TP, FP, and FN. The absence of these metrics makes it challenging to compare the effectiveness of different techniques.
- Several papers in the review mention that their techniques are designed for specific types of sensors or datasets. While this specialization can be beneficial for addressing unique challenges, it limits the generalizability of these techniques to other scenarios, i.e., how these methods can be adapted or extended to handle diverse datasets and sensors.
- Some papers mention using small or medium-sized datasets for training and evaluation. Small-scale datasets may not capture the full complexity and diversity of real-world scenarios. The robustness and scalability of AD techniques should be tested on larger datasets to ensure their practical applicability.
- Many AD techniques mentioned in the review are evaluated using historical data. However, real-time AD is crucial for timely anomaly detection and response in energy systems, real-time monitoring, and decision-making.
- Some papers mentioned specific hardware requirements for their AD techniques. The impact of these requirements on the practicality of implementing these techniques in real-world applications is essential as they may affect the feasibility of the solution.
- Imbalanced datasets, where anomalies are rare compared to normal data, can pose challenges for AD. To overcome this issue, it is vital to consider a technique that addresses imbalanced data or employs a strategy such as oversampling, undersampling, or using different evaluation metrics.
- While the review discusses various energy systems (power generation, energy consumption, PV devices, and solar sensors), the AD techniques discussed for one system do not always apply to other energy systems.
- Many energy systems involve sensitive data. The literature does not extensively address AD techniques' privacy and data security concerns. Ensuring the privacy of data while performing AD is a critical aspect.
- Not all studies discussed the models' availability as open-source software, while access to code and pre-trained models can facilitate further research and adoption in practical applications.
- The review discussed AD techniques' performance but did not adequately explain the real-world implementation challenges in actual energy systems.

- Some papers discuss the impact of external factors, such as weather conditions, on AD, but not all. Also, it is essential to understand the sensitivity of AD techniques to external factors and how they can be accounted for in real-world applications.
- Only a few studies consider the scalability for large-scale energy systems. It is essential to know the performance of these techniques as the data size and the system's complexity increase.

The literature emphasized the need for a tailored approach considering the specific characteristics of the data and problem, underscored the potential of deep learning techniques, and highlighted the challenges related to data annotations and generalization in real-world scenarios. Additionally, it showed some limitations where further research and development are needed to enhance the applicability and robustness of these techniques in real-world scenarios. This research focused on the proposed hybrid model approaches based on rigorous assessments and testing of existing technologies and algorithms to improve and integrate the most relevant concepts into one framework that solves the problem of determining solar data correction.

The existing K.A.CARE's data validation system has the following limitations:

- Lack of a single, systematic approach to validate the correctness of collected data. It relies on manual quality checks and various disconnected tools and processes.
- The quality assurance process in the existing system requires significant human intervention, and it depends heavily on individual expertise. This can be time-consuming and error-prone.

Therefore, this research aims to address and improve upon the shortcomings of the existing system by proposing a design and implementation for an Anomaly Detection (AD) Framework with a decision support system that considers the following elements:

1. **Early Detection of Data Anomalies:** Detecting data anomalies in solar radiation measurements at an early stage using systematic methods for identifying the status of operating sensors based on the correctness of data.
2. **Automated Data Processing:** Automating the data processing steps, including standardization, handling missing data, and incorporating records related to cleaning and maintenance. This reduces the need for manual data manipulation.
3. **Multivariate Time Series Analysis:** Considering the properties of individual time series and exploring relationships and correlations between different time series. This ensures a more comprehensive analysis of data.
4. **Integration of Various Factors:** Incorporating factors such as seasonality, sensor cleaning procedures, calibration, and defining normal and abnormal events. This enables a holistic approach to anomaly detection.
5. **Machine Learning and Deep Learning:** Using ML (e.g., SVM) and DL techniques (e.g., feature extraction, defining normality, and obtaining a score of anomalies along with the stacking fusion) to classify anomaly sources from different data sources (including images) and types (big data). This improves the accuracy and efficiency of anomaly detection.
6. **Human Expert Involvement:** While automated techniques are employed, the framework acknowledges the importance of domain experts' involvement in reviewing and inputting rules and labels and analyzing the causes of abnormal behaviors during the data training phase. This enhances the accuracy of anomaly detection.
7. **Privacy Preservation:** Conducting careful validation to ensure data privacy while implementing the AD Framework, considering the importance of privacy preservation.
8. **Operator Alerts:** Offering output includes identifying anomalies and describing their causes, error levels, and alerts for operators.
9. **Data availability:** The proposed AD framework has advantages over other techniques in the literature for its historical and real-time big datasets availability for learning, training, and testing.

These improvements aim to enhance the accuracy, efficiency, and reliability of data validation in the context of solar radiation measurements, ultimately leading to better data quality and decision-making. Furthermore, the proposed AD framework must be assessed regarding factors like accuracy, classification speed, tolerance to noise and missing values, and learning speed concerning the dimensionality and the number of examples. The following section will introduce the AD framework proposal to solve this research question.

Proposed Anomaly Detection Framework

The proposed AD framework addresses the problem of validating solar radiation data correctness to detect sensors' issues at early stages, determine the cause, and correct the issue. Figure 1 describes the framework's different components and steps.

The big time-series energy data, including Solar radiation measurements, meteorological parameters, station location, elevation, coordinates, images, and sky conditions, are monitored by 46 solar ground stations.

Then, the multivariate time series is retrieved by the Measurement and Instrumentation Datacenter (MIDC), where the collected data storage and quality plots development occur. The next step will be preparing the data for further analysis and processing work. Using special algorithms, the data, e.g., time, will be standardized in a particular format like day/mo/year/ho/min/sec. Simultaneously, missing data will be handled by imputation or replaced with (NAN) as value. Moreover, other records related to cleaning and maintenance are to be incorporated with the data to help in the analysis step. In multivariate time series analysis, the properties of individual time series and the relationships and correlations between different time series should be considered.

Next, identify factors to be considered, such as seasonality (fluctuation, cloud, dust), sensor cleaning procedure, and calibration, as well as defining normal and abnormal events and types of anomalies such as equipment failure, device misalignment, shading, aerosol optical depth AOD. Later, ML (e.g., SVM) and DL strategies are combined to classify information from different sources and types (data, images) to classify anomaly sources and order them by their score for analysis. Although supervised methods are not preferred as they require labeled datasets to learn the abnormalities, these techniques allow the detection of various types of anomalies if they are predefined by human experts using training data. This demonstrates why, at this step (DMS), the framework required domain expert involvement to review and input rules and analyze the causes of the abnormal behaviors detected; FMEA analysis can be used too. Additionally, SVM is advantageous in generating an optimal hyperplane (decision plane) that maximizes the margin between two classes, using support vectors, and does not have hardware demands.

The next component of the framework utilizes multiple DL techniques for AD at 1-min data resolution to suit the different types of data, e.g., segmentation annotation and/or stacking methods. Ultimately, the expected output is a flag indicating the anomaly's cause, a description of the error level (high or low), and an alert for the operator(s). At this stage, careful validation must be conducted to evaluate performance, assess implementation complexity, and investigate privacy preservation.

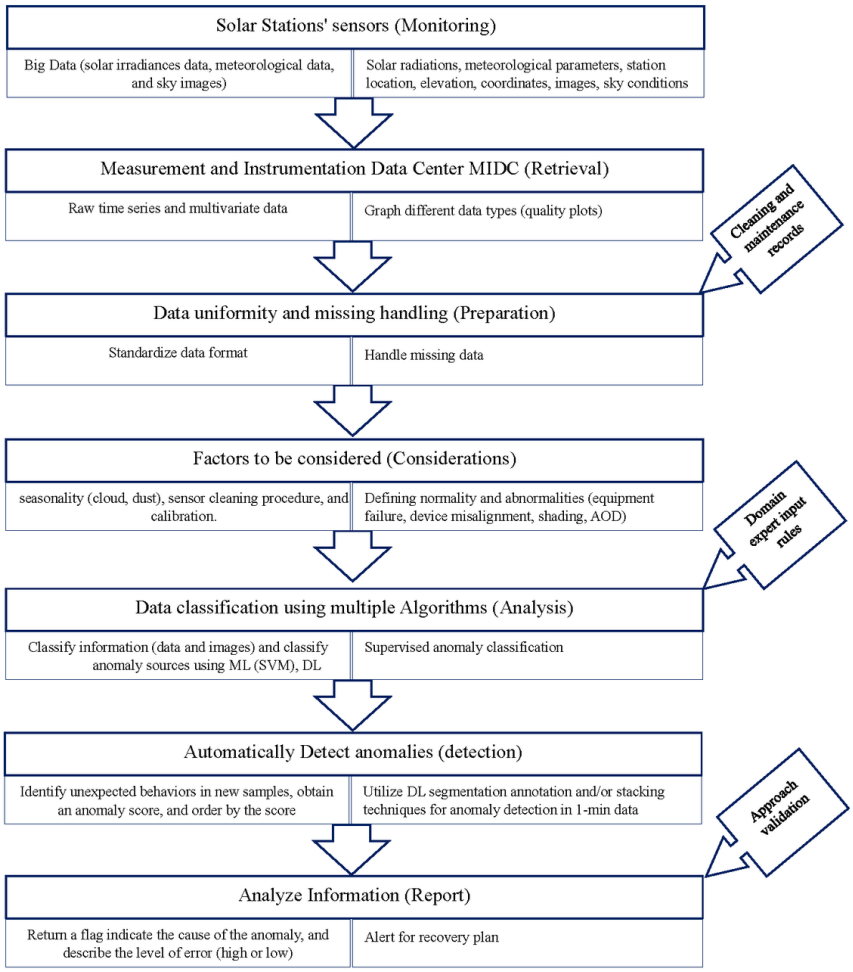
METHODOLOGY

The future roadmap to achieve the framework's objective involves several components and steps.

Data Collection and Monitoring

- The framework begins with collecting big time-series energy data from 46 solar ground stations. This data includes solar radiation measurements, meteorological parameters, station location, elevation, coordinates, images, and sky conditions.
- Continuous monitoring of this data is essential to detect any anomalies or irregularities.

Figure 1. The proposed framework for anomaly detection in renewable energy big data



Data Storage and Quality Control

- The collected data is then sent to the Measurement and Instrumentation Datacenter (MIDC).
- At MIDC, data storage and quality control processes take place. This involves checking for data completeness, consistency, and accuracy.
- Quality plots are developed to visualize data quality and identify potential issues.

Data Preparation

- The next step involves preparing the data for further analysis. This includes standardizing the time format to a specific format (e.g., day/month/year/hour/minute/second).
- Handling missing data through imputation or marking as (NAN) is crucial for ensuring data integrity.
- Additional records related to cleaning and maintenance activities are essential for the subsequent analysis.

Multivariate Time Series Analysis

- In this phase, the properties of individual time series are studied.
- Relationships and correlations between different time series are examined to understand the complex interactions within the data.

Identification of Factors and Anomalies

- Data quality factors are identified, including seasonality, sensor cleaning procedures, calibration, and various sources of anomalies such as equipment failure, device misalignment, shading, and aerosol optical depth (AOD).
- The framework defines what constitutes normal and abnormal events.

Machine Learning and Deep Learning Integration

- Supervised machine learning and deep learning strategies are employed to classify information from different sources and types (data and images) to detect and categorize anomaly sources.
- Supervised methods can detect anomalies for different variables, i.e., a multi-class output model.
- Faulty incidents must be Predefined (labeled) in the datasets. This is one of the strengths of this framework, as there is big historical data that includes will-defined issues and available expertise in the area to help label abnormal data.
- Support Vector Machines (SVM) is chosen as an ML technique for classification for its ability to generate an optimal decision plane that maximizes the margin between two classes.

Domain Expert Involvement

- At the Data Management Stage (DMS), domain experts play a crucial role.
- They review and input rules and analyze the causes of abnormal behaviors detected by the ML/DL models.
- Failure Mode and Effects Analysis (FMEA) analysis can also be employed.

Deep Learning for 1-Minute Resolution Data

- Multiple DL techniques are utilized for AD, especially for high-resolution (1-minute) data.
- Segmentation, annotation, and stacking techniques are applied to different data types.

Output and Alerts

- The expected output of the framework includes a flag indicating the anomaly's cause, a description of the error level (high or low), and an alert for the operator(s).
- Careful validation is essential at this stage to assess the framework's performance, implementation complexity, and considerations for privacy preservation.

Results Evaluation

Evaluating the framework involves assessing its effectiveness, efficiency, and reliability. Since data correctness is crucial for country projects, several evaluation methods and metrics must be considered, for offline (historical data) and online (real-time), to validate its performance in different scenarios. The evaluation metrics include the following:

- **Benchmarking:** Compare the performance of the proposed framework with existing anomaly detection system.
- **Accuracy:** Measure the overall accuracy of anomaly detection to ensure that true anomalies are correctly identified.
- **Precision and Recall:** Calculate precision (positive predictive value) and recall (sensitivity) to understand the trade-off between correctly identifying anomalies and minimizing false alarms.
- **F1-Score:** Uses a single metric that combines precision and recall to assess the system's overall performance.
- **Receiver Operating Characteristic (ROC) Curve:** Draw the ROC curve and compute the Area Under the Curve (AUC) to assess the model's ability to discriminate between normal and abnormal events.
- **Scalability:** Evaluate how well the framework performs as the volume of data or the number of monitoring stations increases.
- **Data Imputation Accuracy:** If missing data imputation is performed, assess the accuracy of imputed values compared to ground truth.
- **Cross-Validation:** Employ cross-validation techniques to assess how well the model generalizes to new data.
- **FMEA Analysis:** Use Failure Modes and Effects Analysis to assess the framework's ability to identify and prioritize critical anomalies.
- **Data Privacy:** Ensure that privacy preservation mechanisms are effective, and that sensitive data is adequately protected.
- **Evaluate:** How well the alerts contribute to detection and resolving sensor issues and anomalies early.

Iterative Improvement

- As the collected solar data is time series (offline) and is continuously collected (online), It is essential to consider providing a constant online learning system that supports AD to improve systems quality based on feedback and evolving data patterns.
- Regular updates to anomaly detection algorithms and domain-specific rules are necessary to adapt to changing conditions.

The future roadmap for this proposed AD framework involves:

- Refining each step,
- Improving anomaly detection accuracy, and
- Ensuring that the system can operate effectively in real-time.

Additionally, ongoing collaboration between data scientists, domain experts, and operators will be crucial for the framework's success. Furthermore, there is a potential to generalize the approach to detect and classify anomalies for general time series data, which is not necessarily generated from solar sensors.

CONCLUSION

This study reviews anomaly detection (AD) techniques in renewable energy (RE) and the importance of solar irradiance in the energy field. The review paper identified the relevant studies of big data AD in the energy field and categorized the related concepts such as supervised, unsupervised, and semi-supervised tasks.

In Addition, we highlighted the ability of DL algorithms to validate data correctness and prepare them appropriately to enable effective AD. The literature presented existing techniques and algorithms to improve solar radiation detection efficiency while considering weather conditions and other factors, such as predefined error instances in the time-series data. There are numerous methods for estimating solar radiation; some are frequently used (ANN, ARIMA, CNN methods), others are used more frequently (SVM, SVR, k-mean), and few are rarely used (boosting, regression tree, RF). In some cases, one is superior to the other, so the ANN and ARIMA methods are equivalent in prediction quality under certain variability conditions. However, the flexibility of ANN as a universal nonlinear approximation makes it preferable to classical ARIMA. In general, the accuracy of these methods is determined by the quality of the training data. Due to the promising results and similar error statistics, the three promising methods are SVM, regression trees, and RF. The errors reported in the literature may have more to do with the methods' implementation than the methods themselves.

DL is widely used in AD for big data, particularly solar radiation data. However, different models' performance depends on many factors related to application and data type. For instance, building an AD for solar data depends on factors, including location, elevation, and weather conditions. Therefore, we proposed an AD framework based on DL methodologies comprising a set of algorithms to address the problem of how to validate solar data correctness concerning the sensors' issues that influence data correctness. The future direction will focus on implementing and evaluating the proposed framework.

DATA SETS

The data supporting this study's findings are available online only at monthly resolution from King Abdullah City for Atomic and Renewable Energy (K.A.CARE). Restrictions apply to the availability of these data at the 1-min resolution, which was used under license for this study. Data is available from the corresponding author, Suzan Katamoura, with the permission of K.A.CARE and under the stated terms and conditions.

AUTHOR NOTE

The authors of this publication declare there are no competing interests. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. Funding for this research was covered by the authors of the article. Thank you to Dr. Abdullah AlMusned and Dr. Raed Sherif for comments on a draft of this article.

Correspondence concerning this article should be addressed to Suzan Katamoura, college of Information Systems, King Saud University, Imam road, *Riyadh, Saudi Arabia, katamoura@hotmail.com.*

REFERENCES

AlOtaibi, Z. S., Khonkar, H. I., AlAmoudi, A. O., & Alqahtani, S. H. (2020). Current status and future perspectives for localizing the solar photovoltaic industry in the Kingdom of Saudi Arabia. *Energy Transitions*, *4*(1), 1–9. doi:10.1007/s41825-019-00020-y

Balzategui, J., Eciolaza, L., & Maestro-Watson, D. (2021). Anomaly detection and automatic labeling for solar cell quality inspection based on Generative Adversarial Network. *Sensors (Basel)*, *21*(13), 4361. doi:10.3390/s21134361 PMID:34202285

Barhoumi, E. M., Okonkwo, P. C., Zghaibeh, M., Belgacem, I. B., Alkanhal, T. A., Abo-Khalil, A. G., & Tlili, I. (2020). Renewable energy resources and workforce case study Saudi Arabia: Review and recommendations. *Journal of Thermal Analysis and Calorimetry*, *141*(1), 221–230. doi:10.1007/s10973-019-09189-2

Benninger, M., Hofmann, M., & Liebschner, M. (2019). *Online Monitoring System for Photovoltaic Systems Using Anomaly Detection with Machine Learning* NEIS 2019; Conference on Sustainable Energy Supply and Energy Storage Systems, Germany. ISBN: 978-3-8007-5152-5.

Blaga, R., Sabadus, A., Stefu, N., Dughir, C., Paulescu, M., & Badescu, V. (2019). A current perspective on the accuracy of incoming solar energy forecasting. *Progress in Energy and Combustion Science*, *70*, 119–144. . doi:10.1016/j.pecs.2018.10.003

Branco, P., Gonçalves, F., & Costa, A. (2020). Tailored Algorithms for Anomaly Detection in Photovoltaic Systems. *Energies*, *13*(1), 225. doi:10.3390/en13010225

Buerhop-Lutz, C., Deitsch, S., Maier, A., Gallwitz, F., Berger, S., Doll, B., Hauch, J., Camus, C., & Brabec, C. J. (2018). *A Benchmark for Visual Identification of Defective Solar Cells in Electroluminescence Imagery* 35th European Photovoltaic Solar Energy Conference and Exhibition (EU PVSEC), Brussels, Belgium. doi:10.4229/35thEUPVSEC20182018-5CV.3.15

Chen, X., Karin, T., & Jain, A. (2022). Automated defect identification in electroluminescence images of solar modules. *Solar Energy*, *242*, 20–29. doi:10.1016/j.solener.2022.06.031

Correa-Jullian, C., Cardemil, J., Droguett, E., & Behzad, M. (2019). *Assessment of Deep Learning Algorithms for Fault Diagnosis of Solar Thermal Systems*. SWC. 10.18086/swc.2019.08.03

De Benedetti, M., Leonardi, F., Messina, F., Santoro, C., & Vasilakos, A. (2018). Anomaly detection and predictive maintenance for photovoltaic systems. *Neurocomputing*, *310*, 59–68. doi:10.1016/j.neucom.2018.05.017

Deitsch, S., Christlein, V., Berger, S., Buerhop-Lutz, C., Maier, A., Gallwitz, F., & Riess, C. (2019). Automatic classification of defective photovoltaic module cells in electroluminescence images. *Solar Energy*, *185*, 455–468. doi:10.1016/j.solener.2019.02.067

Espinar, B., Aznarte, J.-L., Girard, R., Moussa, A., & Kariniotakis, G. (2010). *Photovoltaic Forecasting: A state of the art*. 5th European PV-Hybrid and Mini-Grid Conference, Tarragona, Spain. https://minesparis-psl.hal.science/hal-00771465.

Faber, K., Żurek, D., Pietroń, M., & Piętak, K. (2021). *Ensemble neuroevolution based approach for multivariate time series anomaly detection*, *23*(11). .10.3390/e23111466

Feng, M., Bashir, N., Shenoy, P., Irwin, D., & Kosanovic, B. (2020). *SunDown: Model-driven Per-Panel Solar Anomaly Detection for Residential Arrays*, /arXiv.2005.12181, https://lass.cs.umass.edu/papers/pdf/compass20-sundown.pdf10.1145/3378393.3402257

Fuyuki, T., Kondo, H., Yamazaki, T., Takahashi, Y., & Uraoka, Y. (2005). Photographic surveying of minority carrier diffusion length in polycrystalline silicon solar cells by electroluminescence. *Applied Physics Letters*, *86*(26), 262108. doi:10.1063/1.1978979

Ghofrani, M., & Alolayan, M. (2018). *Time Series and Renewable Energy Forecasting*. InTech Open. 10.5772/intechopen.70845

Harimi, M., & Shayegan Fard, M. (2020). A Method for Anomaly Detection in Big Data based on Support Vector Machine. *International Journal of Information and Communication Technology Research*, *11*, 42–48. https://www.researchgate.net/publication/346659737_A_Method_for_Anomaly_Detection_in_Big_Data_based_on_Support_Vector_Machine

Harrou, F., Abdelkader, D., Taghezouit, B., & Sun, Y. (2019). An unsupervised monitoring procedure for detecting anomalies in photovoltaic systems using a one-class Support Vector Machine. *Solar Energy*, *179*, 48–58. doi:10.1016/j.solener.2018.12.045

Hempelmann, S., Feng, L., Basoglu, C., Behrens, G., Diehl, M., Friedrich, W., Brandt, S., & Pfeil, T. (2020). *Evaluation of unsupervised anomaly detection approaches on photovoltaic monitoring data 2671-2674,* 10.1109/PVSC45281.2020.9300481

Himeur, Y., Ghanem, K., Alsalemi, A., Bensaali, F., & Amira, A. (2021). Artificial intelligence based anomaly detection of energy consumption in buildings: A review, current trends and new perspectives. *Applied Energy*, *287*, 116601. doi:10.1016/j.apenergy.2021.116601

Hu, B. (2012). *Solar Panel Anomaly Detection and Classification* [Master, University of Waterloo, UWSpace]. http://hdl.handle.net/10012/6731

Huang, L., Kang, J., Wan, M., Fang, L., Zhang, C., & Zhaoliang, Z. (2021). Solar Radiation Prediction Using Different Machine Learning Algorithms and Implications for Extreme Climate Events. *Frontiers in Earth Science (Lausanne)*, *9*, 596860. doi:10.3389/feart.2021.596860

Ibrahim, M., Alsheikh, A., Awaysheh, F. M., & Alshehri, M. D. (2022). Machine Learning Schemes for Anomaly Detection in Solar Power Plants. *Energies*, *15*(3), 1082. doi:10.3390/en15031082

Iyengar, S., Lee, S., Sheldon, D., & Shenoy, P. (2018). *SolarClique: Detecting Anomalies in Residential Solar Arrays*. 10.1145/3209811.3209860

Kardi, M., AlSkaif, T., & Tekinerdogan, B. Catalao, Joao P.S., (2021). Anomaly Detection in Electricity Consumption Data using Deep Learning. *IEEE International Conference on Environment and Electrical Engineering*. IEEE. https://edepot.wur.nl/564487

Khan, Z. A., Hussain, T., Haq, I. U., Ullah, F. U. M., & Baik, S. W. (2022). Towards efficient and effective renewable energy prediction via deep learning. *Energy Reports*, *8*, 10230–10243. doi:10.1016/j.egyr.2022.08.009

Klinsuwan, T., Ratiphaphongthon, W., Wangkeeree, R., Wangkeeree, R., & Sirisamphanwong, C. (2023). Evaluation of Machine Learning Algorithms for Supervised Anomaly Detection and Comparison between Static and Dynamic Thresholds in Photovoltaic Systems. *Energies*, *16*(4), 1947. https://www.mdpi.com/1996-1073/16/4/1947. doi:10.3390/en16041947

Kosek, An., & Gehrke, O. (2016). *Ensemble regression model-based anomaly detection for cyber-physical intrusion detection in smart grids.* 16th annual IEEE Electrical Power and Energy Conference, Ottawa, Canada. https://www.epec2016.ieee.ca/

Liu, L. (2018). *Big Data Analytics for PV Systems Real-time Monitoring* [Master, University of Oslo]. Oslo, Norway. http://hdl.handle.net/10852/62817

Malakar, S., Goswami, S., Ganguli, B., Chakrabarti, A., Roy, S. S., Boopathi, K., & Rangaraj, A. G. (2021). Designing a long short-term network for short-term forecasting of global horizontal irradiance. *SN Applied Sciences*, *3*(4), 477. doi:10.1007/s42452-021-04421-x

Mayr, M., Hoffmann, M., Maier, A., & Christlein, V. (2020). Weakly Supervised Segmentation of Cracks on Solar Cells using Normalized Lp Norm, *EEE International Conference on Image Processing (ICIP)*, Taipei, Taiwan, doi:10.1109/ICIP.2019.8803116

Mochizuki, T., Kim, C., Yoshita, M., Mitchell, J., Lin, Z., Chen, S., Takato, H., Kanemitsu, Y., & Akiyama, H. (2016). Solar-cell radiance standard for absolute electroluminescence measurements and open-circuit voltage mapping of silicon solar modules. *Journal of Applied Physics*, *119*(3), 034501. doi:10.1063/1.4940159

Mulongo, J., Atemkeng, M., Ansah-Narh, T., Rockefeller, R., Nguegnang, G. M., & Garuti, M. A. (2020). Anomaly Detection in Power Generation Plants Using Machine Learning and Neural Networks. *Applied Artificial Intelligence*, *34*(1), 64–79. doi:10.1080/08839514.2019.1691839

Natarajan, K., Kumar, B. P., & Kumar, V. (2020). Fault Detection of Solar PV system using SVM and Thermal Image Processing. *International Journal of Renewable Energy Research*, *10*, 967–977.

Oprea, S., Adela, B., Puican, F., & Radu, I. (2021). Anomaly Detection with Machine Learning Algorithms and Big Data in Electricity Consumption. *Sustainability (Basel)*, *13*(19), 10963. doi:10.3390/su131910963

Otamendi, U., Martinez, I., Olaizola, I., & Quartulli, M. (2022). *A scalable framework for annotating photovoltaic cell defects in electroluminescence images, 19*(9), 9361-9369. .10.1109/TII.2022.3228680

Pereira, J., & Silveira, M. (2018). *Unsupervised Anomaly Detection in Energy Time Series Data Using Variational Recurrent Autoencoders with Attention*, 1275-1282, https://doi.org/10.1109/ICMLA.2018.00207

Pol, A. (2020). *Machine Learning Anomaly Detection Applications to Compact Muon Solenoid Data Quality Monitoring*. [Thesis, University of Paris-Sacla]. https://cds.cern.ch/record/2790963/files/TS2020_033_2.pdf

Rahimzadeh, R., Brox, T., & Steinmetz, J. (2022). *State-of-the-art Deep Learning Anomaly Detection Method for Analyzing Electroluminescence Images of Solar Cells*. Conference: Silicon PV 2022, Konstanz. .10.1063/5.0141116

Rahman, M., Khan, I., & Alameh, K. (2021). Potential measurement techniques for photovoltaic module failure diagnosis: A review. *Renewable & Sustainable Energy Reviews*, *151*, 1–17. doi:10.1016/j.rser.2021.111532

Roka, S., Diwakar, M., & Karanwal, S. (2022). *A Review in Anomalies Detection Using Deep Learning. Proceedings of Third International Conference on Sustainable Computing*, Singapore. doi:10.1007/978-981-16-4538-9_33

Rossi, B., Chren, S., Buhnova, B., & Pitner, T. (2016). *Anomaly Detection in Smart Grid Data: An Experience Report*. SMC. 10.1109/SMC.2016.7844583

Sanz-Bobi, M. A., San, R. A. M., De Marcos, A., & Bada, M. (2012). *Intelligent system for a remote diagnosis of a photovoltaic solar power plant* in Journal of Physics Conference Series. 25th International Congress on Condition Monitoring and Diagnostic Engineering (COMADEM 2012) June 2012, Huddersfield, UK. Vol. 364, doi:10.1088/1742-6596/364/1/012119

Sarker, I. H. (2021). Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. *SN Computer Science*, *2*(6), 420. doi:10.1007/s42979-021-00815-1 PMID:34426802

Sharif, M. H., Gupta, K., Mohammed, M., & Jiwani, N. (2022). ANOMALY DETECTION IN TIME SERIES USING DEEP LEARNING. *International Journal of Engineering Applied Sciences and Technology*, *7*(6), 296–305. https://www.researchgate.net/publication/365748435_ANOMALY_DETECTION_IN_TIME_SERIES_USING_DEEP_LEARNING

Soni, U. R., & Sun, X. (2021). *Anomaly detection on big data system logs using deep learning*. https://hdl.handle.net/10211.3/218647

Srinivasan, R., Wang, L., & Bulleid, J. L. (2020). Machine learning-based climate time series anomaly detection using convolutional neural networks. *Weather and Climate*, *40*(1), 16–31. doi:10.2307/27031377

Tang, W., Yang, Q., Xiong, K., & Yan, W. (2020). Deep learning based automatic defect identification of photovoltaic module using electroluminescence images. *Solar Energy*, *201*, 453–460. doi:10.1016/j.solener.2020.03.049

The International Trade Administration. (2021). *Saudi Arabia Renewable Energy (trade.gov)*. Trade. https://ww**w.tra**de.gov/country-commercial-guides/saudi-arabia-power

Toshniwal, A., Mahesh, K., & Jayashree, R. (2020). *Overview of Anomaly Detection techniques in Machine Learning*. 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 808-815. IEEE. doi:10.1109/I-SMAC49090.2020.9243329

Tsai, C. W., Yang, C. W., Hsu, F. L., Tang, H. M., Fan, N. C., & Lin, C. Y. (2020). Anomaly Detection Mechanism for Solar Generation using Semi-supervision Learning Model. 2020 Indo – Taiwan 2nd International Conference on Computing, Analytics and Networks (Indo-Taiwan ICAN), 2nd International Conference on Computing, Analytics and Networks (Indo-Taiwan ICAN), Taiwan. doi:10.1109/Indo-TaiwanICAN48429.2020.9181310

Vinisha, F. A., & Sujihelen, L. (2022). Study on Missing Values and Outlier Detection in Concurrence with Data Quality Enhancement for Efficient Data Processing *4th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, Tirunelveli, India. doi:10.1109/ICSSIT53264.2022.9716355

Voyant, C., Notton, G., Kalogirou, S., Nivet, M.-L., Paoli, C., Motte, F., & Fouilloy, A. (2017). Machine learning methods for solar radiation forecasting: A review. *Renewable Energy*, *105*, 569–582. doi:10.1016/j.renene.2016.12.095

Yang, R., Du, B., Duan, P., He, Y., Wang, H., He, Y., & Zhang, K. (2020). Electromagnetic Induction Heating and Image Fusion of Silicon Photovoltaic Cell Electrothermography and Electroluminescence. *IEEE Transactions on Industrial Informatics*, *16*(7), 4413–4422. doi:10.1109/TII.2019.2922680

Zell, E., Gasim, S., Wilcox, S., Katamoura, S., Stoffel, T., Shibli, H., Engel-Cox, J., & Al Subie, M. (2015). Assessment of solar radiation resources in Saudi Arabia. *Solar Energy*, *119*, 422–438. doi:10.1016/j.solener.2015.06.031

Zhao, Y., Liu, Q., Li, D., Kang, D., Lv, Q., & Shang, L. (2019). Hierarchical Anomaly Detection and Multimodal Classification in Large-Scale Photovoltaic Systems. *IEEE Transactions on Sustainable Energy*, *10*(3), 1351–1361. doi:10.1109/TSTE.2018.2867009

Zhou, H., Liu, Q., Yan, K., & Du, Y. (2021). Deep Learning Enhanced Solar Energy Forecasting with AI-Driven IoT. *Wireless Communications and Mobile Computing*, *2021*, 1–11. doi:10.1155/2021/9249387

Suzan Katamoura is a Ph.D. student at King Saud University, Riyadh, Saudi Arabia, at the College of Computer and Information Systems. Her research focuses on Electronic Business, Knowledge Management Systems, data mining, anomaly detection, and data quality.

Mehmet Sabih Aksoy is a Professor of Artificial Intelligence, in College of Computer and Information Sciences, King Saud University, Riyadh Saudi Arabia. His area of interest includes AI and Knowledge-Based Systems, Electronic Business, Machine Learning, Inductive Learning, Expert Systems, Artificial Neural Networks, and Data Mining.