Machine Learning for Accelerating the Discovery of High-performance Low-cost Solar Cells

Satyam Bhatti

Department of Electronics and Nanoscale Engineering
University of Glasgow

This dissertation is submitted for the degree of
Master of Philosophy

James Watt School of Engineering May 2023

I would like to dedicate this thesis to my loving parents Mrs Balbinder Kaur and Mr Har Bhupinder Kumar.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements.

Satyam Bhatti

May 2023

Acknowledgements

I would like to express my heartfelt gratitude to my supervisors, Dr Sajjad Hussain and Dr Rami Ghannam, for their unwavering support, guidance, and encouragement throughout my MPhil journey. Their insightful feedback, rigorous critique, and constructive suggestions have been invaluable to the development of this thesis.

I am grateful to my research partners, Dr Ruy Sebastain Bonilla (University of Oxford) and Dr Bruno Michel (IBM Research), for their collaboration and contributions to this research. I would like to thank my colleagues and classmates, Habib Ullah Manzoor and Ahsan Raza Khan for their stimulating discussions, thoughtful comments, and support.

I am grateful to the academic and administrative staff, Julia Deans and Heather Lambie who have assisted me with administrative and logistical matters throughout my MPhil. Their support has been greatly appreciated.

Lastly, I would like to thank my family and friends for their unwavering support, encouragement, and love. Their constant belief in me has been a source of strength and inspiration throughout my MPhil journey.

Abstract

Solar energy has the potential to enhance the operation of electronic devices profoundly and is the solution to the most important challenge facing humanity today. Such devices primarily rely on rechargeable batteries to satisfy their energy needs. However, since photovoltaic (PV) technology is a mature and reliable method for converting the Sun's vast energy into electricity, innovation in developing new materials and solar cell architectures is becoming more important to increase the penetration of PV technologies in wearable and IoT applications. Moreover, artificial intelligence (AI) is touted to be a game changer in energy harvesting. The thesis aims to optimize solar cell performance using various computational methods, from solar irradiance and solar architecture to cost analysis of the PV system. The thesis explores the PV cell architectures that can be used for optimized cost/efficiency trade-offs. In addition, machine learning (ML) algorithms are incorporated to develop reconfigurable PV cells based on switchable complementary metal-oxide-semiconductor (CMOS) addressable switches, such that the output power can be optimized for different light patterns and shading.

The first part of the thesis presents a critical literature review of a range of ML techniques applied for estimating solar irradiance, followed by a review on accurately predicting the levelized cost of electricity (LCOE) and return on investment (ROI) of a PV system and lastly, presents a systematic review (SR) on the discovery of solar cells. Furthermore, the literature review consists of a thorough systematic review that reveals that ML techniques can speed up the discovery of new solar cell materials and architectures. The review covers a broad range of ML techniques that focus on producing low-cost solar cells. Additionally, a new classification method is introduced based on data synthesis, ML algorithms, optimization, and fabrication process. The review finds that Gaussian Process Regression (GPR) ML technique with Bayesian Optimization (BO) is the most promising method for designing low-cost organic solar cell architecture. Therefore, the first part of the thesis critically evaluates the existing ML techniques and guides researchers in discovering solar cells using ML techniques. The literature review also discusses the recent research work done for predicting solar irradiance and evaluating the LCOE and ROI of the PV system using various time-series forecasting techniques under ML algorithms.

Secondly, the thesis proposes an ML algorithm for accurately predicting solar irradiance using the wireless sensor network (WSN) relying on batteries that need constant replacement and are hazardous waste. Therefore, WSNs with solar energy harvesters that scavenge energy from the Sun are proposed as an alternative solution. Consequently, the ML algorithms that enable WSN nodes to accurately predict the amount of solar irradiance are presented so that the node can intelligently manage its energy. The nodes use the panel's energy to power its internal electronic components, such as the processor and transmitter, and charge its battery. Accordingly, this helps the node access an exact amount of solar irradiance predictions to plan its energy utilization more efficiently, thereby adjusting the operation schedule depending on the expected solar energy availability. The ML models were based on historical weather datasets from California, USA, and Delhi, India, from 2010 to 2020. In addition, the process of data pre-processing, followed by feature engineering, identification of outliers, and grid search to determine the most optimized ML model, is evaluated. Compared with the linear regression (LR) model, the support vector regression (SVR) model showed accurate solar irradiance forecasting. Moreover, from the predicted output calculated results, it was also found that the models with time duration of 1 year and 1 month have much better forecasting results than 10 years and 1 week, with both root square mean error (RMSE) and mean absolute error (MAE) less than 7% for California, USA.

Consecutively, the third part of the thesis evaluates the parameter LCOE using demographic variables. Moreover, LCOE facilitates economic decisions and quantitative comparisons between energy generation technologies. Previous methods for calculating the LCOE were based on fixed singular input values that do not capture the uncertainty associated with determining the financial feasibility of a PV project. Instead, a dynamic model that considers important demographic, energy, and policy data that include interest rates, inflation rates, and energy yield is proposed. All these parameters will undoubtedly vary during a PV system's lifetime and help determine a more accurate LCOE value. Furthermore, comparisons between different ML algorithms revealed that the ARIMA model gave an accuracy of 93.8% for predicting the consumer price of electricity. Moreover, the proposed model with two case studies from the United States and the Philippines is evaluated in detail. Results from these case studies revealed that LCOE values for the State of California could be almost 30% different (5.03 ¢/kWh for singular values in comparison to 7.09¢/kWh using our ML model), which can distort the risk or economic feasibility of a PV power plant. Additionally, the ML model predicts the ROI of a grid-connected PV plant in the Philippines to be 5.37 years instead of 4.23 years which gives a clear indication to the client for making an accurate estimation for the cost analysis of a PV plant.

Research Publications

- **Journal Articles**

1. **S. Bhatti**, H. U. Manzoor, B. Michel, R. Bonilla, R. Abrams, A. Zoha, S. Hussain, and R. Ghannam, "Revolutionizing Low Cost Solar Cells with Machine Learning: A Comprehensive Review of Optimization Techniques," Wiley Advanced Energy and Sustainability Research: Accepted.

2. **S. Bhatti**, A. R. Khan, A. Zoha, S. Hussain, and R. Ghannam, "A machine learning framework for predicting the lcoe of pv systems using demographic, energy and policy data," IEEE Transactions on Energy Markets, Policy and Regulation: Accepted subject to minor revisions.

- **Conference Proceedings**

1. **S. Bhatti**, A. R. Khan, S. Hussain, and R. Ghannam, "Predicting renewable energy resources using machine learning for wireless sensor networks," in 2022 29th IEEE International Conference on Electronics, Circuits and Systems (ICECS), IEEE, 2022, pp. 1–4, https://ieeexplore.ieee.org/abstract/document/9970851.

Table of contents

List of figures

List of tables

Nomenclature

Acronyms / Abbreviations

AACODS Authority, Accuracy, Coverage, Objectivity, Date, Significance

AI Artificial Intelligence

ARIMA Autoregressive Integrated Moving Average

BO Bayesian Optimization

CMOS Complementary metal-oxide-semiconductor

CNN Convolutional Neural Network

CPE Consumer Price of Electricity

DFT Density functional theory

DSSC Dye-Sensitized Solar Cells

ESCLATE Experiment Specification, Capture and Laboratory Autonomous Technology

ETL Electron Transport Layer

EXTR Extra Tree Regressor

GBRT Gradient Boost Regression Trees

GPR Gaussian Process Regression

HOIP Hybrid Organic Inorganic Perovskites

HOMO Highest Occupied Molecular Orbit

HTL Hole Transport Layer

ITC Inverse Temperature Crystallization

KNN k-Nearest neighbours

LCOE Levelized Cost of Electricity

LR Linear regression

LSTM Long Short-Term Memory Model

LUMO Lowest Unoccupied Molecular Orbit

ML Machine Learning

MLP Multi layer Perceptron

OSC Organic Solar Cells

PCE Power Conversion Efficiency

PRISMA Preferred Reporting Items for Systematic Review and Meta-Analysis

PSC Perovskite Solar Cells

PV Photovoltaics

RF Random Forest

ROI Return on Investment

SVM Support Vector Machine

SVR Support Vector Regression

TDM Transient Decay Measurements

VASP Vienna Ab-initio simulation package

Chapter 1

Introduction

1.1 Background

Current miniature portable and implantable devices rely on batteries that need replacement and are hazardous to patients [1–3]. Surgical removal is required when replacing batteries in implantable devices, which may be inconvenient for patients [4, 5]. Moreover, implantable biomedical devices are often powered using wires, which may cause discomfort, skin infections, and other hazards to patients [6]. The key issues with implanting batteries include metal poisoning for patients due to battery degradation, thus leading to malfunction in generating signals and the damage of electronic circuits [7].

Due to their high energy density, scavenging solar energy using photovoltaic (PV) cells has emerged as a potential and feasible solution to power miniature portable devices [8, 9]. In general, the architecture of these solar cells can be designed as regular, inverted, mesoporous or planar structures. Furthermore, solar cells combine various materials to enable efficient photon absorption, electron transport, and electron extraction to an external circuit. This means there are vast opportunities for discovering solar cell materials and architectures. In fact, solar cell fabrication techniques involve optimizing different coating materials, thermal annealing conditions, encapsulation methods, etc., which often takes place in the research laboratory [10].

However, despite their benefits, these harvesters still suffer from poor efficiency, weak stability, rigidity, and a relatively high cost [11]. Promising PV technologies that aim to overcome issues with rigidity and high cost include Perovskite Solar Cells (PSC), Organic Solar Cells (OSC), and Dye-Sensitized Solar Cells (DSSCs) [12]. Despite rapid progress in the PSC and OSC field, the stability and efficiency of these low-cost, thin-film solar cells are still poor due to the effects of moisture and temperature [13]. Consequently, machine

learning (ML) and artificial intelligence (AI) can be used to improve the performance and accelerate the discovery of these low-cost solar cells [14].

From the systems perspective, ML algorithms can also help develop reconfigurable PV cells based on switchable CMOS addressable switches [5] by developing an optimization method of switchable CMOS addressable switches, followed by prediction of PV cell behavior and real-time control of the operation of reconfigurable PV cells in the real-time. In addition, conjugation is a key characteristic of organic materials, which are frequently used in such devices, and it plays a crucial part in low-cost solar cells. Conjugated polymers or tiny molecules with alternate single and double bonds frequently make up the organic components in solar cells. For the effective conversion of solar energy, conjugation enables the organic materials to absorb light in the visible region of the spectrum. An exciton, which is an excited state produced when a conjugated substance absorbs light, can be split into electrons and holes to produce an electrical current. The performance of low-cost solar technology depends highly on conjugated materials' capacity to transport these electrons and holes through the device effectively [15].

Innovation in developing new low-cost solar cells is needed, which can be achieved with the help of experimentally validated finite element modelling using software tools such as Sentaurus TCAD. However, this is a time-consuming effort, and leveraging the power of AI can be a game changer in discovering new materials and fabrication techniques to help expedite the process of selection, design, and optimization [16]. Furthermore, the distribution of electron density in the energy levels of materials used in solar cell architecture, known as the HOMO (Highest Occupied Molecular Orbital) and LUMO (Lowest Unoccupied Molecular Orbital) pattern, is a crucial factor that affects the solar cell's efficiency in capturing photons and producing electrical energy. Matching the HOMO and LUMO levels of different materials used in the cell is a significant challenge in solar cell design to optimize charge separation efficiency and minimize recombination, which results in energy loss and decreased efficiency [17, 18].

To investigate the characteristics of charge carriers (electron and hole) in solar cell materials, researchers use a method called Transient Decays Measurements (TDM) analysis. This analysis involves monitoring the decay rate of photo-generated carriers over time following a transient pulse of light. When the material absorbs sunlight, it creates electron-hole pairs that produce a photocurrent in the solar cell. The TDM analysis tracks the time it takes for the photocurrent, which is related to the recombination of electron-hole pairs, to decay. During the recombination process, charge carriers combine and cancel each other out, causing energy loss and reducing the efficiency of the solar cell [19].

Furthermore, in the literature, ML relates to the development and ability of the model to learn to adapt, forecast, and predict the independent variables [20]. ML algorithms consist of 3 types: Supervised learning, Unsupervised learning, and Reinforcement learning [21]. The supervised ML approach takes the input data from the user to learn from past experiences and, accordingly, trains the model [22]. However, the unsupervised ML train model depends upon the real-time data generated and outputs depending on the information given by the user. In contrast, reinforcement learning is the subset of ML that enables an AI-driven system (also known as an agent) to learn by performing tasks and receiving feedback from its trials and errors [23]. Herein, various ML techniques are discussed in-depth to find an optimized structure for solar cells [24].

Examples of ML techniques reported in the literature include linear regression, logistic regression, k-nearest neighbours (KNN), random forest (RF), etc., [25, 26] however; every problem requires a unique ML algorithm [27]. Every algorithm has unique abilities and data requirements. For instance, linear regression would not be very helpful due to nonlinear relations in solar cells. For logistic regression, an assumption that factors are independent of each other is made, which might not be the case in solar cells. Similarly, KNN aims to locate the nearest neighbours with the best possible value. So, the use of ML in optimizing solar cells depends upon the type of experiment, optimizing variables, and data type.

Since the fabrication of OSCs is cheap, most experimental work is carried out via trial and error, which does not guarantee the best performance [28]. Instead, researchers are now turning their attention to data-driven techniques for material design and discovery [29]. ML is one of the vital data-driven techniques that is rising to prominence in discovering new solar cells, forecasting electrical characteristics, and performance prediction without any experimentation [30, 31]. ML uses algorithms to visualize and analyze data that has several advantages over traditional programming techniques [32]. Chapter 3 systematically reviews the different ML algorithms used to find an optimized structure of a low-cost solar cell. The output power can be optimized for different light conditions and shading depending on the positioning of the solar cells [33]. The integration of ML methods for designing low-cost solar cells is thoroughly discussed and, consecutively, explores the literature on using different ML techniques for the advanced discovery of solar cells.

1.2 Aim and Objectives

1.2.1 Aim

The thesis aims to present a critical literature review, followed by implementing the most recent ML algorithms that could be applied to predict the amount of solar irradiance on a WSN device and, lastly, to develop an ML framework for estimating the LCOE and ROI of a PV system.

1.2.2 Objectives of the thesis

The following are the key objectives of the MPhil thesis:

1. To present a literature review of the ML techniques for predicting solar irradiance and estimating a PV system's LCOE and ROI. Also, to present an SR of research articles discussing the discovery of low-cost solar cells.

2. To propose an ML model capable of accurately predicting the amount of solar irradiance from WSNs with solar energy harvesters that scavenged energy from the sun.

3. To propose a dynamic ML model that accurately estimates the LCOE and ROI of a PV system considering important demographic variables and energy policy data and includes interest rates, inflation rates, and the energy yield and validate the ML models with two case studies.

1.3 Motivation of the thesis

The traditional batteries used in miniaturized portable and implantable devices need frequent change over time and pose a risk to patients. When batteries in implantable devices need to be changed, surgical removal is necessary, which may be uncomfortable for patients. Furthermore, wires used to power implantable biomedical devices frequently put patients at risk for discomfort, skin infections, and other problems. The main problems with battery implants are metal poisoning in patients brought on by battery deterioration, which results in signal generation problems and electronic circuit damage. Solar energy harvesting methods are thus one of the most important battery substitutes. Additionally, WSN nodes also rely on dangerous batteries that require regular replacement. So, solar energy harvesters on WSNs scavenge energy from the Sun. The main problem with these harvesters is that solar power is sporadic. Furthermore, the earlier approaches for figuring out the LCOE relied on

fixed, singular input values that failed to account for the uncertainty of whether a PV project would be financially feasible. Therefore, to address these problems, a dynamic model that incorporates crucial demographic, energy, and policy data, such as interest rates, inflation rates, and energy yield, is proposed in this thesis.

1.4 Contributions to the thesis

Following are the key contributions that are made to the thesis:

1. Conducted a literature review of research articles involving the prediction of solar irradiance and estimation of LCOE and ROI of a PV system.
2. Systematically review the literature on low-cost solar cells using ML techniques to investigate the techniques used for optimizing solar cells with the help of ML.
3. Proposed ML algorithms to accurately predict the amount of solar irradiance so that node can intelligently manage its own energy and determine the most optimized ML model for the prediction.
4. Proposed an ML model that accurately estimates the LCOE and ROI of a PV system using a dynamic model that takes into account important demographic, energy, and policy data that includes interest rates, inflation rates, and energy yield.
5. Validated the proposed model with two case studies from the United States and the Philippines to compare various ML models to measure the accuracy and loss function.

1.5 Organisation of the thesis

The organization of the thesis is as follows. Chapter 1 of the thesis discusses the background of the study under consideration, followed by aims, objectives, motivations, contributions, and the overall organization of the thesis. A critical literature review for solar irradiance and LCOE of PV systems, along with the adopted methodology in reviewing the literature for an SR, is critically discussed in Chapter 2, which also includes the overall results of the SR in response to the research questions. Consecutively, Chapter 3 showcases ML algorithms for accurately predicting renewable energy resources and the amount of solar irradiance that enables WSN nodes. The chapter includes an introduction, a system model, a model training process, results, and discussions of the proposed ML model. In addition to this, Chapter 3 also discusses areas of further study, future outlook, recommendations, and open

research issues. Followed by Chapter 4 presents an ML framework for accurately predicting the LCOE and ROI of the PV system. Chapter 4 incorporates an introduction, literature review, methodology, data explanation, results, and discussions of the various ML models. Lastly, Chapter 5 is the conclusions and future work of the thesis, where the conclusion, open questions, and future outlook of the thesis are thoroughly discussed.

Chapter 2

Literature Review: Machine learning for solar cells and PV systems

2.1 Introduction

During the past 5 years, there has been a surge in the use of ML and AI techniques for designing new solar cells [34, 35]. In this chapter, previously published review papers are reviewed on this field using ML techniques, and further, discuss their limitations as well as the contributions that this chapter provides to the literature.

Qiuling *et al.* [36] reviewed the ML techniques for only perovskite materials design and discovery. However, their review lacks a comprehensive comparison of ML techniques for other low-cost solar cells, such as organic, inorganic, hybrid, and DSSCs. Additionally, Hannes *et al.* [37] discussed the challenges of ambient hybrid solar cells for IoT devices, while the paper presented by Hannes *et al.* [38] reveals the study on solar cell cracks using statistical parameters of electroluminescent images using ML. However, both studies presented limited ML algorithms to explore solar cell electrical characteristics.

Furthermore, Yongjie *et al.* [39] reviewed recent advances in computational chemistry for OSC discovery and mentioned the DFT, time-dependent DFT, all atomic molecular dynamics, and coarse-grained molecular dynamics. Although their review covered OSCs, it lacked the ML techniques to expedite the process. Next, Florian *et al.* [40] reviewed the literature on designing light-harvesting devices using ML, but the review was limited to only OSCs. Likewise, a review paper presented by Sheng *et al.* [41] covered only ML optimization of PCSs. The studies presented by Anton *et al.*[42], Min-Hsuan *et al.* [43], and Cagla *et al.* [16] explored ML approaches to discover solar cell performance analysis. However, a major

drawback in these studies was that limited ML approaches were discussed and did not involve the scope for optimization as well as the fabrication of solar cells in the real environment.

Therefore, based on the above, state-of-the-art review articles on ML for solar cell discovery focused mainly on a single ML technique with a set of input data. This chapter initially presents a thorough literature review of research papers that incorporate ML techniques for predicting solar irradiance and then reviews the previously published research articles consisting of ML algorithms for the estimation of LCOE and ROI of a PV system. Lastly, the chapter aims to systematically review the range of ML techniques for developing solar cells. These ML techniques include the procedure to pre-process the input data, various ML algorithms, optimization, and fabrication of the solar cell in a real environment. In this context, the review goes beyond existing literature, showcasing how various ML techniques can accelerate the discovery of high-performance, low-cost solar cells.

2.2 ML for Predicting Solar Irradiance

In order to optimize the use of renewable energy sources in WSNs, which are necessary for long-term network functioning, Sharma et al. [44] suggested an ML-based solution. The authors provide a model that forecasts future patterns of energy availability and use using previous meteorological information and sensor readings. Then, in order to reduce the energy deficit in the network, they utilize a reinforcement learning algorithm to assign energy sources based on the projected energy availability and consumption patterns. Simulations are used to assess the suggested strategy, and the results show that it can successfully balance energy supply and demand while ensuring long-term network operation.

In another study, Sharma et al. [45] evaluated a method for predicting the production of daily global solar irradiance. The amount of solar energy that can be produced in a specific place on a given day is predicted by the authors using historical meteorological data and satellite photos. To create predictive models, they combine feature engineering and ML techniques like random forest and gradient boosting. The suggested method is tested on a dataset of solar irradiance measurements obtained from an Indian solar power plant, and the findings demonstrate that it can predict daily solar irradiance generation with high accuracy. According to the authors, this strategy can be used to increase the effectiveness and dependability of solar power generation systems, which can have major positive effects on the environment and the economy.

Furthermore, to estimate the power production of solar panels in a WSN utilized in precision agriculture, Dhillon et al. [46] suggested a neural network-based solar energy forecast model. The amount of solar energy that can be produced by the solar panels is

predicted by the authors using historical weather data and satellite pictures. The predictive model is then constructed using a neural network technique, specifically a feed-forward neural network with backpropagation. The results demonstrate that the suggested method can estimate solar power generation accurately using data gathered from a WSN installed in a precision agriculture area. According to the authors, this method can be used to reduce the amount of energy that WSNs in precision agriculture utilize.

Additionally, the power production was estimated for the solar panels in a WSN utilized in precision agriculture, Ghuman et al. [47] suggested a neural network-based solar energy forecast model. The amount of solar energy that can be produced by the solar panels is predicted by the authors using historical weather data and satellite pictures. The predictive model is then constructed using a neural network technique, specifically a feed-forward neural network with backpropagation. The results demonstrate that the suggested method can estimate solar power generation accurately using data gathered from a WSN installed in a precision agriculture area. According to the authors, this method can be used to reduce the amount of energy that WSNs in precision agriculture utilize.

2.3 ML for LCOE of PV System

Numerous examples in the literature describe the statistical and probabilistic models for calculating the Levelized cost of electricity (LCOE) and energy return on investment (ROI). For example, K. Branker et al. [48] argued that there is a lack of understanding of the calculations involving assumptions and justifications for the estimation of LCOE, thus proving that poor assumptions lead to contradictory results for the calculations of energy return on investment of a PV system. In their paper, they calculated the LCOE to reduce the assumptions-based model and represent a more accurate one; however, the study was limited to singular inputs for calculating the LCOE. In addition, a more detailed calculation of LCOE by Chul-Yong Lee et al. [49] represented a stochastic model for calculating LCOE for solar PV systems installed in the Philippines. Their results depicted that for a commercial solar panel, the LCOE ranged from a minimum of 10 ¢/kWh to 18 ¢/kWh, and they did a sensitivity analysis to validate their results. However, the study lacks the optimized value of the LCOE and only discussed a range of possible LCOE values in their model.

Another study for a utility-based system installed in IESCO, Pakistan, conducted by Ahsan et al. [50] showed an analysis for forecasting day ahead load demand using the Auto-Regressive (AR), Moving Average (MA) and Auto-Regressive Integrated Moving Average (ARIMA) model for the statistical modeling for the load demand. In addition, they did the comparative analysis using the ML techniques like Artificial Neural Networks (ANN) and

Bagged Regression Tree (BRT). However, their results for forecasting the load demand using various ML techniques are precise but lack the estimation of LCOE and hence, the energy return of investment of the model. Furthermore, Geissmann *et al.* [51] showed a probabilistic approach for computing the LCOE of a nuclear plant and a gas power project. Furthermore, they implemented a Monte Carlo simulation to determine the dependency of singular input parameters on the model's final results. However, their study used singular inputs and lacked dependency on demographic variables. Further, Georgitsioti *et al.* [52] discussed the formula used to calculate the LCOE based on singular values for domestic PV systems in the UK, and the financial benefits that can be gained from a domestic PV system under the "Feed in Tariff (FiT)" PV supporting policy in the UK.

2.4 ML for the Discovery of Solar Cells

This section of the chapter presents a systematic review (SR) of the discovery of low-cost solar cells using ML techniques. Firstly, the methodology of SR elaborates the process of shortlisting the research articles depending on the data-driven approach, ML techniques used, optimization processes, and fabrication techniques.

2.4.1 SR Methodology

This section discusses the research objectives and the methodology in collecting and synthesizing the SR on ML algorithms for designing and fabricating low-cost, high-performance solar cells. The four key objectives of the SR are: (1) To review the range of ML techniques for designing low-cost solar cells using historical data.; (2) To identify the ML techniques used specifically for discovering new PV materials.; (3) From a device perspective, identify the specific ML and optimization techniques used for designing efficient solar cell architectures. ; and (4) To identify ML algorithms specifically used for fabricating low-cost PV cells from the circuits and systems perspective.

Figure 2.1 maps the four research objectives and the process of shortlisting the research articles. Initially, the chapter focuses on extracting and pre-processing the historical data, followed by discovering new materials and optimising solar cells. Lastly, the research articles that discuss the integration of ML for fabricating solar cells are reviewed. Accordingly, for the SR, the research objectives are defined to target a set of questions that are the need for the study. Additionally, a set of research articles are shortlisted using the search engines available on Google for extracting the recent research articles published in this domain.

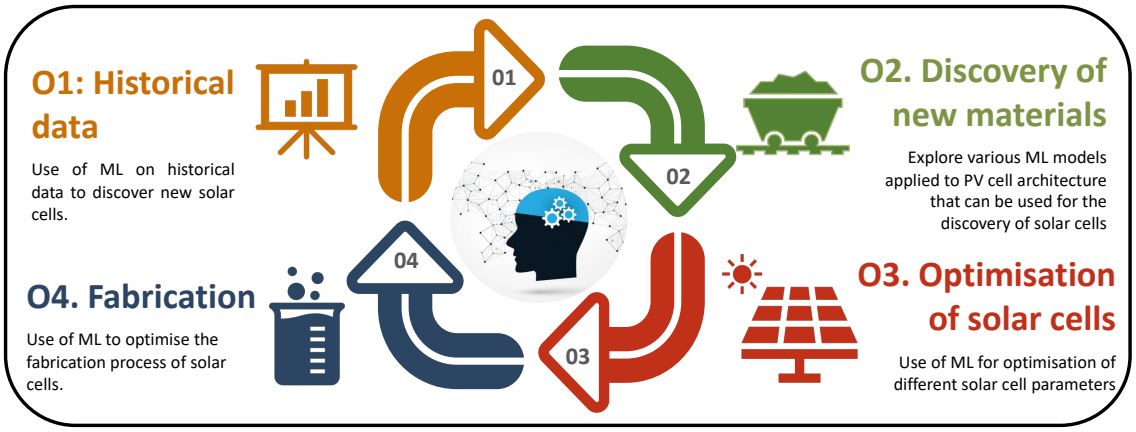

Fig. 2.1 The research objectives for the systematic review.

The SR aims to answer four research questions: (1) What are the data-driven approaches for designing low-cost, high-performance solar cells? ; (2) How can ML algorithms facilitate the discovery of new low-cost solar cell materials? ; (3) What optimisation techniques are used for designing an efficient, low-cost solar cell architecture? ; and (4) What ML algorithms are used for fabricating low-cost solar cells from a circuits and systems perspective?

A proper review protocol is instigated for structuring the SR, and the following are the perquisites of the adopted analogy. This section discusses the search strategy, inclusion criteria, exclusion criteria, and screening mechanisms for selecting relevant research papers. The review considered the latest research articles from major publishing houses that include IET, Science Direct, Nature, AIP, Wiley, IEEE Explorer, IoP science, ACS publications, and MDPI. The search also included non-pre-reviewed articles from arXiv. Thus, a critical appraisal is performed using the AACODS (Authority, Accuracy, Coverage, Objectivity, Date, Significance) checklist as an evaluation and critical appraisal tool of grey literature (publications and research created by groups not affiliated with conventional academic or commercial publishing institutions).

The search began with queering all the repositories with different research items. The keywords were defined such as "Machine Learning", "Data-driven approach", "PV cell architecture", "Solar cells", "Low-cost", "Optimization" and "fabrication" for extracting the research articles. Articles were scanned based on their title and abstract as well as a full-text read of the publications. In addition, search strings are developed using Boolean operators (AND, OR) to connect these keywords.

Inclusion Criteria of SR

The following are the parameters used in the inclusion criteria. (1) Included only English-language articles involving the data-driven approaches of designing solar cells using ML techniques and were pertinent to the study issues such as poor data quantity and data quality.; (2) Included the pertinent articles facilitating the discovery of only low-cost solar cells using ML methods before determining their eligibility.; (3) Included comparative studies involving the optimization and robustness of solar cells designed from ML services.; and (4) Targeted only articles that discussed ML for solar cells, solar cell optimization, and publications on ML integration on solar cells.

Exclusion Criteria of SR

The following is a list of the exclusion criteria for shortlisting the research papers based on the research objectives and targeted research questions. (1) Research articles published in languages other than English.; (2) Research papers that are not available in full text.; (3) Editorials, survey reviews, abstracts, and brief papers involving secondary studies are excluded.; (4) Articles that did not address the integration of ML approaches with solar cells and the ones that involved the expensive manufacturing of solar cells.; and (5) The research articles published before 2018 were also excluded due to the unavailability of quality input data that resulted in poor implementation of ML techniques.

2.4.2 Results and Analysis of SR

In this chapter, the shortlisted research articles are discussed, and how they are aligned with the research objectives and questions. Figure 2.2 shows the workflow of the planning (data extraction and data pre-processing), training (applying various ML techniques and comparing the model's accuracy), testing (optimization), and execution (fabricating solar cells in the laboratory) for discovering new solar cell architectures. As previously mentioned, the review focuses on low-cost solar cells such as PSCs, OSCs, and hybrids.

- ***Data Extraction for Solar Cells***

Solar cells are typically designed with specific objectives, such as reliability, affordability, efficiency, and stability. To predict the structure of low-cost solar cells, research is ongoing to gather and analyze data from previous solar cell fabrication experiments in real-world environments. The quantity and quality of the extracted dataset are crucial to the effectiveness of ML algorithms. Based on the literature, larger input datasets generally result in higher

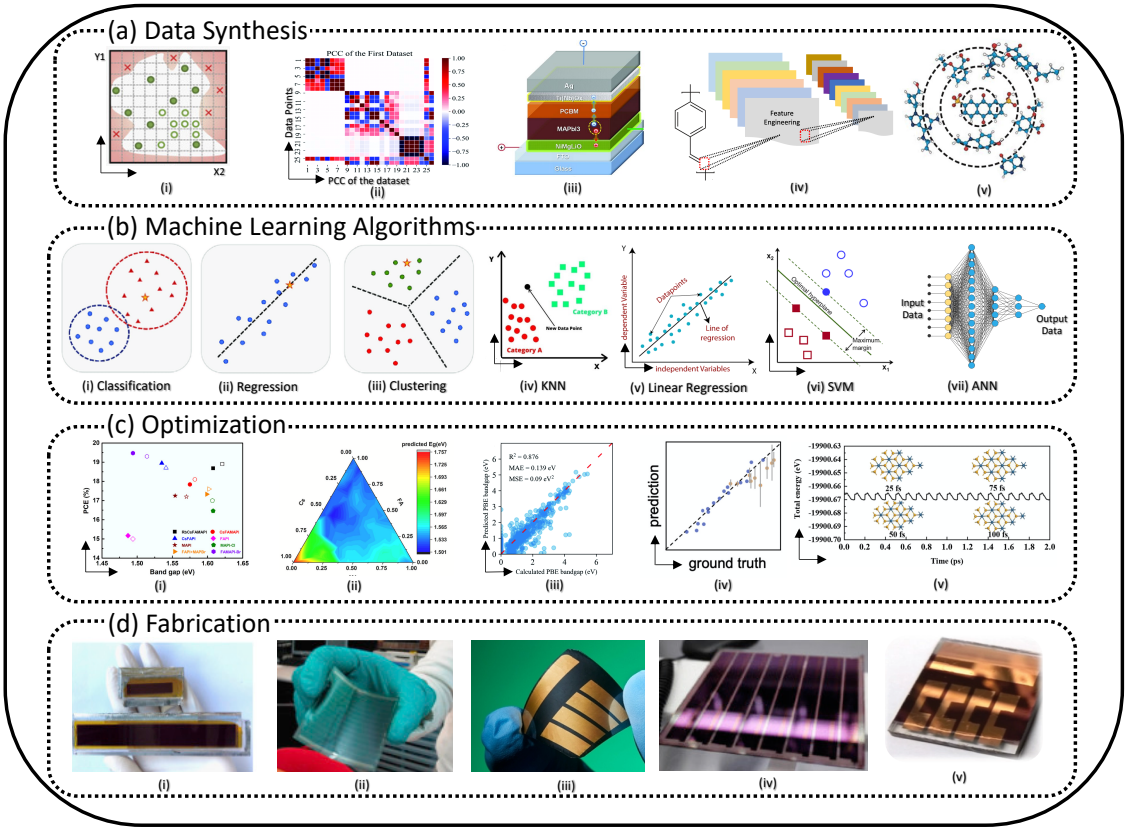

Fig. 2.2 The figure demonstrates the general workflow of the process of discovering low-cost solar cells using ML algorithms.

accuracy and lower functional error values. Consequently, this section focuses on addressing RQ1.

Perovskite Solar Cells

Jino *et al.* [53] investigated how the Gradient Boost Regression Trees (GBRT) ML method [54] can be used for designing Pb-free perovskites. They developed a dataset containing the electronic structures of candidate halide double perovskite. Using the dataset, the GBRT ML model was implemented to predict the values of heat formation and bandgap. Initially, they generated the dataset using two space groups of the crystal structure with 540 hypothetical chemical compounds of $A_2B^{1+}B^{3+}X_6$. Finally, they conducted statistical analysis on the attributes that were chosen to determine design principles for the development of fresh lead-free perovskites.

Moreover, a study presented by Jinxin *et al.* [55] showed how 333 data points from nearly 2000 peer-reviewed papers were used to build ML models for designing PSCs. Their ML models included Linear Regression, KNN, RF and Artificial Neural Networks (ANN) for building two forecasting models, material property characteristics and device performance

prediction. The higher R-value proves that the expected trend is consistent with actual experiments and PSC physics. The highest theoretically computed solar cell efficiency curve depending on the solar spectrum has a bandgap area in the range of 1.15-1.35 eV, and this bandgap region predicts a PCE of above 25%.

Moreover, Felipe *et al.*, [28] demonstrated a new data-driven optimization framework to bridge the mismatch between R&D and industrial production of solar cells. Further, their framework incorporated scalable inference and techno-economic analysis using ML approaches to predict the root cause of the underperformance in PSCs. They also compared traditional R&D optimization vs their proposed total revenue optimization framework using linear, binned and non-linear functions. Consequently, they presented a case study for fabricating 144 PSCs choosing 12 various combinations of dominant processes. In addition, they proposed a surrogate-based black-box model such as Gaussian Process Regression (GPR) and Bayesian Optimization (BO) [56].

In a conference, Maniell *et al.* [57] demonstrated how the optoelectronics properties of PSCs can be predicted using ML methods. A model was developed for testing the bandgap of new different types of PSCs, and the bandgap was capable of predicting the chemical properties and material composition. $C_{Sx}MA_{1-x}PbI_3$, $CsPb(I_xBr_{1-x})_3$ and $MAPb_{1-x}Sn_xI_3$ were the perovskite materials used for testing and resulted in bandgaps ranging from 1.3-2.3 eV. In addition, their study presented a curve showing the predicted PCE values from the ML model vs the actual PCE from fabricated samples. Moreover, another result showed that the predicted value of the fabricated $CsSnI_3$ was 1.15 eV whereas the fabricated sample had a bandgap of 1.25 eV. Lastly, their research article discussed various ML models such as ANN, Random forest algorithm, and Support Vector Regression.

In addition, the robot accelerated discovery and investigation of PSCs were demonstrated by Zhi Li *et al.* [58]. The article presented an automated, high-throughput method for evaluating single crystals of metal halide perovskites based on inverse temperature crystallization (ITC) in order to quickly pinpoint and perfect the conditions for the synthesis of high-quality single crystals. Using 45 organic ammonium cations, a total of 8172 metal halide perovskite synthesis processes were carried out. The screening enhanced the number of metal halide perovskite materials by five times and resulted in designing a new combination of PSCs such as $[C_2H_7N_2][PbI_3]$ and $[C_7H_{16}N_2][PbI_4]$. In addition, to enable experiment generation and data management, they used a software pipeline called ESCALATE (Experiment Specification, Capture and Laboratory Autonomous Technology). Further, their research added 17 new materials (a 400% increase) of metal halide perovskites, which are accessible via ITC. This helped identify conditions that lead to the formation of perovskite single crystals consisting of 19 of 45 target perovskite compositions.

In 2020, Yun *et al.* [30] investigated the ML lattice constants for cubic perovskite A_2XY_6 compounds. Their dataset included a broad spectrum of Fmm group perovskite halides and a total of 79 samples. With lattice constants ranging from 8.109 A to 11.790 A, 79 cubic perovskite compounds were investigated. The ionic radii of [K, Cs, Rb, Tl], [Ge, Mn, Ni, Pd, Pt, Si, Cr, Pd, Ir, Mo, Pb, Re, Se, Ta, Sn, Te, Ti, W, Zr, Ru, Tc, Po, U, Os, Hf], and [F, Cl, Br, I] were among those used as descriptors. The GPR was used for determining the relation between the ionic radii and the lattice constants for cubic perovskites. They used MATLAB for the computational exploration of the model and achieved CC, RMSE, and MAE of 99.72%, 65%, and 0.44%, respectively.

In addition, Chenglong *et al.* [59] presented a two-step ML approach for PSC design, which was based on 2006 PSCs data points taken from peer-reviewed articles published between 2013 and 2020. The authors developed heuristics for high-efficiency PSC and thus, improving PCE dependent on doping of the ETL. The main characteristic of their study was to determine the development of high-performance PCE of PSCs. Their research showed that using SnO_2 and TiO_2 ETLs, mixed-cations perovskites, dimethyl sulfoxide, and dimethylformamide, as well as anti-solvent treatment, led to even higher PCEs. Lastly, they predicted that FA-MA-based PSC with a Cs-doped TiO_2 ETL and a Cs-FA-MA-based PSC with an S-doped SnO_2 ETL were also expected to show PCEs of up to 30.47% and 28.54%.

To expedite the identification of prospective PV cells from 2D perovskites, Hong-Jian *et al.* [60] integrated atomic-level prediction with ML and DFT. Their model implemented a gradient boosting regressor (GBR), a random forest regressor (RF), and an extra tree regressor (EXTR) ML for training a dataset of 2303 perovskite materials. Further, the trained model screened out 4828 materials and also pre-screened using DFT structural relaxation validation from 29,285 artificial perovskites. In fact, a maximum PCE of 30.35% and 26.03% was achieved for (Sr_2VON_3 and Ba_2VON_3).

Likewise, Elif *et al.* [61] predicted the overall performance and bandgap in PSCs. In her analysis, she used eight different PSCs to forecast the bandgap and PCE of perovskites. Initially, they performed the bandgap estimation of perovskites from Tauc plots on a UV-vis spectroscopy using the RF regression ML model with more than one decision tree and experimental approach. Later, they developed a model showing the J-V spectra predicted values for calculating the PCE. Their results showed that perovskites with bandgaps exceeding 0.99 eV could be used to model various new lead halide structure perovskites depending on the accurately predicted value of the bandgap.

Another case study presented by Xia *et al.* [62] combined ML techniques with an efficient forward-inverse method to research $MAS_{nx}Pb1_xI_3$ material and explored high-performance PSCs. With 14 physicochemical parameters and the Sn-Pb ratio as inputs, the E_g model

of $MAS_{nx}Pb1_xI_3$ was first developed for forward analysis, and the asymmetrically bowing relationship between the Sn-Pb ratio and the Eg of OMHP was used. The established NN-based models for PSC performance models showed good predictions for the data points and offered significant insights for PSC devices. Further, for the performance model, a comparison of the prediction model was made with the ML algorithms such as LR, SVR, KNR, RFR, and GBR. In fact, ML models with GBR performed best with values of R2, RMSE, and MAE reaching 0.9172, 0.0386, and 0.0325.

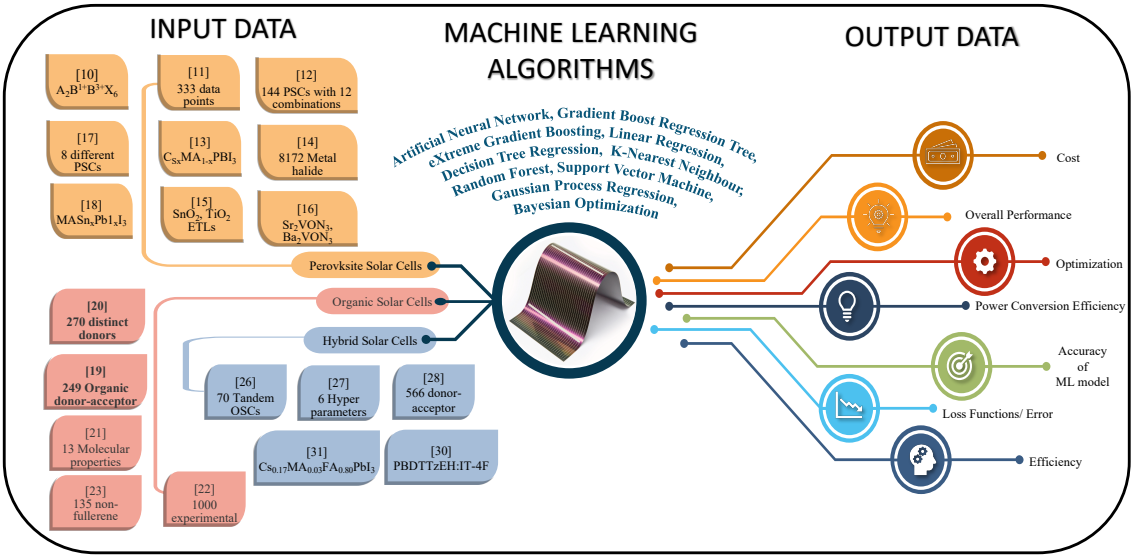

Fig. 2.3 Input data for various materials that are reviewed based on the defined research questions for 3 types of solar cells.

Organic Solar Cells

A rigorous framework involving the classification of the chemical structures in materials discovery was presented by Shinji *et al.*[63]. Further, the dataset of 249 Organic donor-acceptor pairs was computed based on equilibrium geometries and electronic properties such as DFT simulations. Initially, their study discussed predictions using Scharbar's model and resulted in a small energy bandgap of 1.5 eV between the experimental and the computational energy bands. Moreover, they implemented k-NN regression for predicting OSCs characteristics and their PCEs. Finally, the study concluded that k-NN results in correlations of 0.6, which were further improved to 0.7 by implementing non-linear kernel methods.

In addition, Harikrishna *et al.* [64] investigated the PCE of OSCs using ML techniques. They developed a dataset of 280 small molecule OSCs with 270 distinct donors. Firstly, they analyzed the significance of orbitals in the energy conversion process and developed ML models using the characteristics of organic compounds to estimate the PCE for high

throughput virtual screening. In another study, they implemented ML methods to study the correlations between the molecular properties and the device characteristics of an OSC [65]. The authors designed ML methods based on 13 molecular properties as descriptors to predict the three device parameters such as (V_{OC}, J_{SC}, and the fill factor). In addition, the calculations were carried out on Gaussian 09 package for a computational server having Intel Xeon 5115 CPUs. They combined multiple regression trees along with RF and GBRT to incorporate the ML methods. Further, screening of the potential compounds by these models results in high predictive ability (r = 0.7).

Moreover, Daniele *et al.* [66] performed the computer-aided screening of polymers-based OSCs using RF and ANN-based ML supervised-learning models. The dataset involved 1000 experimental characteristics such as PCE, the molecular weight of each organic compound, and other electronic properties. The results showed that the correlation coefficient of ANN was low. However, the RF model achieved better accuracy than the predictive model. Subsequently, Min-Hsuan *et al.* [67] also performed the RFT regression for the analysis of the non-fullerene-based OSCs to predict the overall efficiency of the solar cells. A dataset of 135 non-fullerene acceptor/donor pairs based on OSCs (117 non-fullerene acceptor materials and 30 donor materials) was gathered to examine its electronic properties and device performances. Therefore, their ML model resulted in the highest predictive power by achieving the coefficient of determination (R^2) of 0.85 for the training and 0.80 for testing sets of the ML algorithm.

Furthermore, Xiaoyan *et al.* [68] demonstrated an optimization technique to assess the potential of organic photovoltaic (OPV) materials and solar cell devices for industrial production. They presented an automated characterization of OPV materials, device performance and photostability. The GPR ML technique drove the optimization method with optical absorption characteristics and indicated better prediction accuracies for PV electrical characteristics. Moreover, the efficiency and photostability screening for 100 process conditions were completed in 70 hours. They also proposed a model material system of PM6:Y6, completely automated device fabrication in air resulted in a maximum PCE of 14%.

In one of the latest papers published by Ahmad *et al.* [69], they discuss the implementation of ML to screen small molecule donors for OSCs and molecular descriptors feed ML methods. The co-authors collected a dataset of 340 OSCs devices with donors represented as small molecules while acceptors as fullerenes for the ML-assisted pipeline suitable for small molecule donors for Y6 (an electron acceptor). In addition, they performed ML analysis on an open-source platform called Konstanz Information Miner (KNIME). Further, for training the model, the dataset was divided into training sets, validating sets and external test sets. Also, the descriptors and experimental PCE were used as input to the ML model. They

compared the result depending on various regression techniques, such as RF, LR, SVM and k-NN, for the prediction of PCE. Using data from small donors paired with fullerenes, the SVM model was trained and showed higher prediction ability. The PCE of a few small molecule donors linked with Y6 was predicted using their approach and developed are more than 1000 new small molecule donors. Accordingly, the PCEs were anticipated, and the top 10 applicants with a PCE of over 13% were chosen in their study.

Figure 2.2 demonstrates the general workflow of the process of discovering low-cost solar cells using ML algorithms. The block diagram is divided into four block diagrams, (a) data synthesis, (b) ML algorithms, (c) optimization, and (d) fabrication. For (a) data synthesis, (i) Discusses the data extraction in a statistical form, (ii) Pearson's correlation coefficient matrix, (iii) Solar cell architecture with layer combinations, [70], (iv) data-preprocessing for classification problems [71] and (v) Gradient-based extraction of data. [72] The second block (b) represents the ML algorithms used, (i) Classification, (ii) Regression, (iii) Clustering, [73] (iv) KNN, (v) Linear regression, (vi) SVM, (vii) ANN. [74] The third block (c) discusses the optimization techniques, (i) Bandgap Vs PCE curve, [61] (ii) Ternary contour plots, [75] (iii) Predicted Vs Calculated PCE, (iv) Predicted Vs Ground truth curve, (v) Predicted accuracy of ML model, (vi) Total energy dissipation Vs Time curve. [76] The fourth block discusses the fabricated solar cells. (vi) [77] [78] [79].

Hybrid Solar Cells

Another article presented by Min-Hsuan *et al.* [43] investigated the performance and matching band structure for Tandem OSCs by implementing two ML methods, RF and the SVR. The ML techniques were initially developed using 70 tandem OSCs (37 conventional and 33 inverted tandem OSCs), which were used as the data points. Furthermore, to understand the structure, they calculated Pearson's correlation coefficient. Among the two ML methods, the efficient method for forecasting solar efficiency was the RF Regression having eight electronic features of selection.

Moreover, to address the stability concerns with PSCs, Tianmin *et al.* [80] used a progressive ML algorithm to investigate the impact of input data by providing a reliable and accurate approach for deep mining of the hidden hybrid organic-inorganic solar cells. To predict the electronic bandgaps of HOIP perovskites, they implemented GBR, SVR, and kernel ridge regression (KRR) using material property. The best results from six hyperparameters were chosen. They also used DFT calculations for the chosen HIO perovskites and incorporated them into the Vienna Ab-initio simulation package (VASP). Their results show that the GBR model performs with the highest level of accuracy (R2 = 0.943, MAE = 0.203, MSE = 0.086) when compared to the SVR (R2 = 0.826, MAE = 0.367, MSE = 0.276) and KRR (R2 = 0.819, MAE = 0.387, MSE = 0.288) models.

The effect of enhancing the descriptors using ML prediction for small molecule-based OSCs was discussed by Zhi-Wen *et al.* in his study. [81] The dataset consists of a total of 566 organic donor-acceptor (D/A) pairs found from the literature search, with 513 unique donors and 33 unique acceptors (including C_{60}, $PC_{61}BM$, $PC_{71}BM$, ITIC, IDTBR, IDIC, PDIs, etc.) among the donors. Further, they implemented k-NN, KRR and SVR ML models to predict the PCE of hybrid solar cells. Also, the study examined Pearson's correlation coefficient for all combinations of descriptors, including donor molecules and device parameters.

In another study presented by Yao *et al.*, [82] five different ML algorithms were used and gave 565 donor-acceptor combinations for training the dataset. Furthermore, to implement the material design and donor-acceptor pairs, the screening of non-fullerene in OSCs was performed. They used 565 donor/acceptor (D/A) combinations as training data sets in their study to assess the viability of these ML algorithms for use in directing material design and the screening of D/A pairs. Therefore, the ML techniques RF and BRT offer the best prediction capacities. Additionally, RF and BRT models are screened and estimated to be more than 32 million D/A pairs, respectively. Lastly, six photovoltaic D/A couples are picked and synthesized so that their experimental and predicted PCEs for critical comparison.

In an investigation presented by Kakaraparthi *et al.* [83], the co-authors used the RF model on an experimental dataset consisting of 0.85 correlation coefficient for the ML of non-fullerene and polymer OSCs. Moreover, 200,932 conjugated polymers produced by the combinatorial coupling of acceptor and donor units were screened virtually. Additionally, a number of conjugated polymers centred on benzodithiophene and thiazolothiazole were created, produced, and studied using various alkyl chains in order to assess the efficacy of the ML model. In terms of the selection of alkyl chains, PBDTTzEH: IT-4F demonstrated a PCE of 10.10% and, thus, shows good predictions while using ML techniques.

One of the primary concerns with perovskites is their stability. As a result, Shijing *et al.* [84] demonstrated how to discover the most stable organic-inorganic alloyed perovskites using a sequential learning framework. They introduced a data-fusion approach for estimating Gibbs Free Energy of mixing from DFT and experimentally analyzed degradation using aging tests. Moreover, they applied ML probabilistic constraints in an end-to-end BO approach to combine data from high-throughput degradation testing and first-principle simulations of phase thermodynamics. The results showed that perovskites centered at $Cs_{0.17}MA_{0.03}FA_{0.80}PbI_3$ exhibit low optical change with increased temperature, moisture, and light having more than17-fold stability improvement over $MAPbI_3$ by sampling 1.8% of the discretized $Cs_xMA_yFA_{1xy}PbI_3$ compositional space (*MA*, methylammonium; *FA*, formamidinium; PbI_3, lead halide).

- ***ML to Facilitate the Discovery of Solar Cells***

This section discusses the research articles and peer-reviewed journals related to the discovery of solar cells using ML techniques.

Discovery of Organic Structures

A target-driven approach was provided by Tianmin *et al.* [85] to accelerate the discovery of HOIPs for PV applications from 230808 HOIP candidates. Also, they combined the ML method with DFT calculations. 686 orthorhombic-like HOIPs with the appropriate bandgap were chosen after possible HOIP candidates are subjected to the two criteria of charge neutrality condition and stability condition, followed by an ML screening. In ML screening, ensemble learning was used to forecast the bandgap of 38086 HOIPs candidates using three ML models, including GBR, SVR, and KRR. Finally, 132 stable and non-toxic orthorhombic-like HOIPs (free of Cd, Pb, and Hg) were confirmed by DFT calculations with the proper band gap for solar cells.

Oleksandr *et al.* [86] used ML in-the-loop to learn from the experimental data, suggested experimental parameters to explore, and indicated regions of synthetic parameter space that would permit record-monodispersity PbS quantum dots. Their results show that the technique that produces record-large bandgap (611 nm exciton) PbS nanoparticles with a well-defined excitonic absorption peak (half-width at half-maximum (hwhm) of 145 meV) permits nucleation to triumph overgrowth by adding a growth-slowing precursor (oleylamine). With a hwhm of 55 meV at 950 nm and 24 meV at 1500 nm, respectively, as opposed to the best-published values of 75 and 26 meV, they also improved monodispersity at longer wavelengths.

Double chalcogenide perovskites were investigated in a study presented by Michael *et al.* [87] to find new photovoltaic absorbers that can take the place of CH3NH3PbI3. ML approaches were used to categorize materials as potential photovoltaic absorbers using information from the periodic table, thus avoiding unnecessary computation due to the wide range of possible compounds. On the created data set, a random forest method obtains a cross-validation accuracy of 86.4%. Traditional and statistical approaches are used to identify over 450 potential alternatives, with Ba2AlNbS6, Ba2GaNbS6, Ca2GaNbS6, Sr2InNbS6, and Ba2SnHfS6 emerging as the most promising options when thermodynamic stability, kinetic stability, and optical absorption are taken into account.

Nastaran *et al.* [88] in a study showed that ML techniques used by computationally intensive DFT simulations to quickly and precisely estimate the properties of OPV materials. One-hot descriptors, OPV power conversion efficiency (PCE), open circuit potential (V_{oc}), short circuit density (J_{sc}), highest occupied molecular orbital (HOMO) energy, lowest unoccupied molecular orbital (LUMO) energy, and the HOMO-LUMO gap were all quantified in the study. With a standard error of 0.5 for a percentage of PCE for both the training and test

Table 2.1 Literature Discussing the ML for facilitating the discovery of solar cells

Data Extraction	**Machine Learning Used**	**Solar Cell Architecture**	**Ref**
230808 HOIP, 686 orthorhombic	GBRT, SVR, KRR	Non-toxic orthorhombic-like HOIPs (free of Cd, Pb, Hg)	[81]
Double chalcogenide perovskites, 450 alternatives	ML with DFT simulations	Ba2AlNbS6, Ba2GaNbS6, Ca2GaNbS6, and Ba2SnHfS6	[83]
Organic Photovoltaic Materials, one hot descriptor	Intensive DFT simulations	Design of OPVs pre-screening possible donor and acceptor materials	[84]
21 organic halide salts	Supervised ML and Shapley values	Phenyltriethylammonium iodide (PTEAI)	[85]
3880 unknown spinels	XGBoost method	$CaAl_2O_4$	[86]
227 experimental dataset	RF, XGBoost, LR, k-NN, SVR and MLP	$PC_{61}BM$ and $PC_{71}BM$	[87]
250 OSCs dataset	RF model	ABX_3-type perovskites	[88]
28 million double-perovskite	- - -	17 sodium-, potassium-, and ammonium-based tin-halide perovskites	[89]
N-annulated perylene sensitizers	MLR and QSPR model	C_{281}	[90]
Metal halide perovskites (MHPs)	CNN	0.01 eV, 5 degrees, and 0.01	[91]
10,000 candidates	Quantitative Structure-Property	eight promising organic dyes	[92]
Lead-free halide perovskite material	RF, RR, SVR, and GBRT	Lead-free halide double PSC	[93]
78,400 DHOIPs	Integrating ML techniques	19 promising ones, HSE06 calculations	[94]

sets, the most reliable and predictive models were able to predict PCE (computed by DFT). Their methodology helps to expedite the design of OPVs for use in green energy applications by pre-screening possible donor and acceptor materials.

An ML framework introduced by Noor *et al.* [89] involved optimizing the capping layer of perovskite degradation. They featured 21 organic halide salts, used them as capping layers on (MAPbI3) films, aged them rapidly, and implemented supervised ML and Shapley values to identify factors determining stability. They discovered a correlation between higher MAPbI3 film stability and organic molecules' limited number of hydrogen-bonding donors and tiny topological polar surface area. Phenyltriethylammonium iodide (PTEAI), the best organic halide, successfully increases the stability lifespan of MAPbI3 by 4 2 times over bare MAPbI3 and 1.3 0.3 times over cutting-edge octylammonium bromide (OABr).

Zhilong *et al.* [90] created a target-driven approach that makes use of ML to speed up the ab initio predictions of unidentified spinels from the periodic table. Eight spinels with direct band gaps and thermal stabilities at room temperature are successfully selected out of 3880 unknown spinels using this method ($CaAl_2O_4$, $CaGa_2O_4$, $SnGa_2O_4$, $CaAl_2S_4$, $CaGa_2S_4$, $CaAl_2Se_4$, $CaGa_2Se_4$, $CaAl_2Te_4$). A semiconductor classification model is developed based on the XGBoost method, and it has a strong structure-property link. It has a high prediction accuracy of 91.2% and a low computational cost of a few milliseconds. The suggested target-driven strategy enables the discovery and design of a wide variety of energy materials while cutting the research cycle of spinel screening by about 3.4 years.

The accuracy for predicting the bandgap of an OSC is a vital factor in terms of the characterization of solar cell devices. Accordingly, Yiming *et al.* [75] used ML algorithms to predict the performance of different architectures for the compound ABX_3-type in PSCs. Also, they gathered 227 experimental datasets consisting of the bandgap of perovskites extracted from recently published 1254 publications. For their model, they used ML methods such as RF, XGBoost, LR, k-NN, SVR, and Multilayer perceptron (MLP). Their prediction analysis from ML models showed that B-site metal and the X-site halogen ion have a significant impact on bandgaps of the ABX_3-type perovskites from SHAP explanations.

Muhammad *et al.* [91] did the critical analysis of the small-molecule donors for OSCs such as Fullerene using the ML methods. In order to train the ML model, they used molecular descriptors as an input and consecutively, they implemented a number of ML techniques to measure the best ML algorithm for the desired outcome. The dataset used in the study consists of 250 OSCs having a combination of acceptors and donors as fullerenes ($PC_{61}BM$ and $PC_{71}BM$). They used the platforms like Konstanz Information Miner (KNIME) and Weka platforms to implement the ML model and thus, the Random Forest model resulted

the best predictive model with Pearson's coefficient as 0.93. Lastly, to determine the most efficient materials, the PCE values for the small-molecular donor was predicted.

Discovery of Hybrid Halide Structures

With multiple newly developed, computationally economical, and high-performing (Pearson's correlation coefficient = 0.7-0.8) ML models employing pertinent descriptors, Harikrishna *et al.* [92] carried out high-throughput virtual screening of 10,170 candidate compounds, assembled from 32 distinct building blocks. Furthermore, to create effective molecules, crucial building elements are recognized, and new design principles are implemented. Additionally, 126 candidates are suggested for synthesis and device fabrication with theoretically projected efficiency >8%.

A high-throughput material search scheme based on materials informatics was devised and carried out for PSC materials after Shohei *et al.* [93] explored the existence of viable alternative perovskites. More than 28 million double-perovskite-like compounds were screened using this method. Five well-known organic-inorganic tin-halide perovskites and 17 sodium-, potassium-, and ammonium-based tin-halide perovskites were among the 24 most promising possibilities found. Promising solar cell materials included two perovskites based on transition metals.

Further, Lifei *et al.* [94] constructed N-annulated perylene sensitizers and put forth one goal-directed approach that combined quantum chemical analysis with data mining approaches. By using MLR to build the robust quantitative structure-property relationship (QSPR) model, they were able to identify the key characteristics using a genetic algorithm (GA). The potential dyes were then created using the model's recommendations. The proposed molecules' overall power conversion efficiencies (PCEs) were anticipated by the model to be 15.7%, up 22.0% from reference dyes C_{281}.

For the electrical characteristics of metal halide perovskites (MHPs), which have a billions-range materials design space, Wissam *et al.* [95] employed CNN to create a predictive model. Furthermore, they demonstrated that as compared to simple techniques, a well-designed hierarchical ML strategy offers a higher degree of predictability in terms of MHP features. The bandgap for the MHPs' lattice constants, octahedral angle, and RMSE were all calculated using the hierarchical ML scheme, and the corresponding RMSE values were 0.01 eV, 5 degrees, and 0.01.

Yaping *et al.* [96] combined ML with computational quantum chemistry results in the establishment of an accurate, reliable, and interpretable QSPR model. Using this model, virtual screening as well as the evaluation of synthetic accessibility are carried out to find new effective and synthetically accessible organic dyes for DSSCs. Finally, out of almost

10,000 candidates, eight promising organic dyes with high power conversion efficiency and synthetic accessibility were eliminated.

Moreover, Zongmei *et al.* [97] investigated the discovery of PSC materials via ML stability and calculated the bandgap of lead-free halide perovskite materials. They performed a comparative analysis of four different ML techniques such as the random forest, ridge regression, support vector regression, and the gradient boost regression tree. Among these four ML techniques, XGBoost gave the highest predictive performance i.e. R2:0.9935 and MAE:0.0126 in terms of thermodynamic stability, and accordingly, the random forest gave the highest predictive performance i.e. R2:0.9410 and MAE:0.1492 for bandgap analysis of the lead-free halide double PSCs. Moreover, their study showed an interesting result that XBoost performs best when considering the thermodynamic stability and electronegativity's linear correlation.

By integrating ML techniques, high-throughput screening, and density functional theory, Jialu *et al.* [98] showed the ability to speed up the discovery of double hybrid organic-inorganic perovskites (DHOIPs). In contrast to other studies, the anisotropy of organic cations of DHOIPs was first assessed, and then the properties were predicted using an ML technique using low-level calculations to predict the properties of DHOIPs accurately. From 78,400 DHOIPs, 19 promising ones with suitable bandgaps for solar cells were selected and verified using HSE06 calculations.

John *et al.* [99] investigated the bias, temperature, light, and H2O, O2, and air pressure affected device performance and recovery. They first talked about important studies that assess the 3R cycle's capabilities of perovskites and how ML algorithms may help determine the best values for each operating parameter. They then looked at perovskite dynamics and degradation, highlighting the difficulties in understanding this 3R cycle. Finally, they suggested an ML paradigm with a shared knowledge library for improving long-term performance and forecasting device performance recovery.

Discovery of Solar Cells using NLP

In another study, a framework related to the high-throughput synthesis of the PSCs was discussed with ML image recognition used for automated characterization by Jeffrey *et al.* [100]. Perovskite single-crystal synthesis was carried out at high throughput, and the results were identified using convolutional neural network-based image recognition. Also, they quickly created 96 distinct crystallization environments using a protein drop setter and then examined the crystals. On the other hand, trained a convolutional neural network (CNN) was used to determine if crystals had been produced using a dataset of 7,000 photographs. Then, a larger dataset of 25,000 photos was employed with this classifier. The first synthesis of

$(3-PLA)_2PbCl_4$ was then achieved after they employed ML modeling to predict the ideal conditions for synthesizing a novel perovskite single crystal.

A study presented by Lei *et al.* [101] showed ML techniques based on natural language processing (NLP) to predict the properties of solar cell materials, which were then examined using first-principle calculations. The aim of the study was to reduce the amount of human interaction and enable computers (without supervision) to learn the latent knowledge about solar cell materials depending on the textual data and generate predictions about the composition of solar cells. The first-principles calculations were used to determine the projected material's density of states, UV–vis absorption spectra, as well as band structures in order to assess their suitability for photovoltaic applications. The formula and targeted keywords for solar cells were represented as vectors in the ML process, which facilitated the successful relationship extraction of the materials and their applications. The ML model was validated using first-principles calculations on the unusual solar cell materials included in the list, and the projected candidates, such as As_2O_5 have good electrical and optical characteristics that are suitable for solar cell applications.

- ***ML for Solar Cell Optimization***

The focus of this section is RQ3, which involves examining the optimization techniques used with machine learning algorithms to develop optimized and reconfigurable solar cells. The technical research articles that showed experimental work for implementing the ML algorithms for discovering the optimized solar cells are included.

Moreover, Figure 2.4 displays multiple layered internal architectures of solar cells and the necessary chemical components for creating reconfigurable solar cells. Specifically, Figure 2.4 (a) depicts the perovskite's chemical structure with a carbon composition, whereas Figure 2.4 (b) shows the arrangement of the chemical components in a solar cell and Figur 2.4 (c) shows the various layers of a solar cell that have been sliced for clarity in depicting the solar cell architecture. Finally, Figure 2.4 (d) showcases the outer layer of a solar cell, including Ag, BCP, PCBM, Perovskite, Poly-TPD, ITO, and glass.

Donor/Acceptor Ratio for Higher PCE

Most scientific advancements in the field of materials have been produced experimentally, frequently using one variable at a time testing. However, neither are the properties of materials-based systems straightforward nor related [102]. Authors in [103], claim that the optimization of OSCs has a high level of complexity due to the high complexity and interconnectivity of different components. Changing one component can have an unforeseen impact on other components. Hence ML can play a vital role in the optimization process of

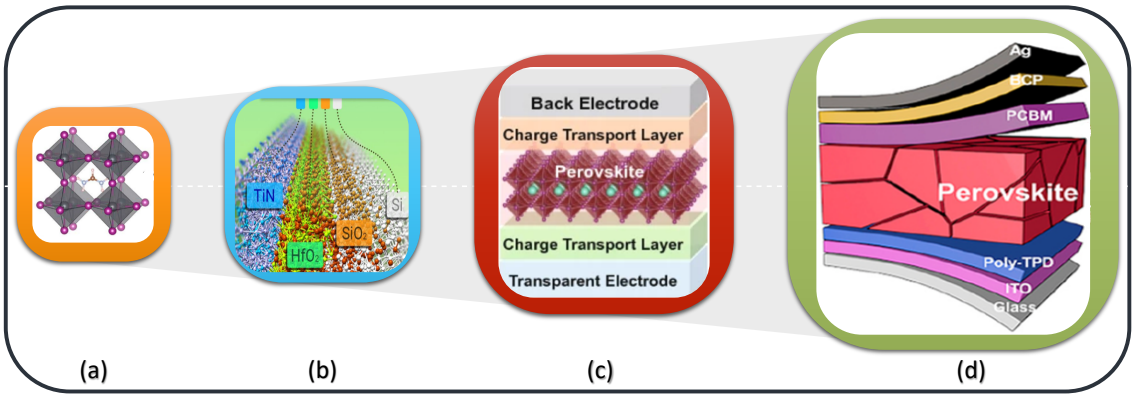

Fig. 2.4 The figure displays multiple layered internal architectures of solar cells and the necessary chemical components for creating reconfigurable solar cells.

OSCs. They used $PDCTBT : PC_{71}$ solar cell and observed the effect of donor/acceptor ratio, total concentration, spin speed, and additive volume on PCE(%). The authors applied SVM using the radial basis function. They conducted two sets of experiments, where they used optimized results of the first experiment in the second experiment and found a significant increase in PCE of fabricated devices [104]. In the first set of the experiment, only three out of fifteen devices were above the threshold (PCE 6.3%); however, in the second, all thirteen devices produced PCE above the threshold.

Conductivity Optimization of Solar Cells

SVM regression was used in [72] for the optimization of $p-CZS/n-si$, $p-CZS/p^{+}n-Si$ heterogeneous solar cells. SVM was implemented with a radial-based function using Scikit-learn [105] in python. They used ten-fold cross-validation to tackle the problem of over-fitting. They predicted the figure of merit (FOM) from film conductivity and optical transmission in desired transmission range. Optimization results show that FOM has increased from 14.8μ to 173μ. Furthermore, current density has increased from 11.8 to 17.9 mA/cm^2 for $p-CZS/n-si$ solar cells and from 13.8 to $18.0 mA/cm^2$ $p-CZS/p^{+}n-Si$ for solar cell. The authors claimed their approach is valid for any general application to any material synthesis process with multiple parameters [106].

Selection of Donor/Acceptor Pairs

From 2010 to 2017, 320 organic donor and acceptor pairs (hetero-junction solar cells) were reported in the literature. These 320 donors and acceptors can make 19912 combinations. Authors in [107] applied distanced-based ML techniques KNN and SVM to optimize PCE.

They have provided a list of unexplored donor and acceptor combinations that can be helpful in the future in fabricating highly efficient solar cells. The use of back propagation neural network, deep neural network, SVM, and the random forest is reported in [108] to predict highly efficient OSCs. The data set contained 1719 realistic donor materials of OSCs. The authors used images, ASCII strings, and fingerprints as input, and out that fingerprints with 1000 bits can provide higher conversion efficiency. The authors also proposed ten new materials.

Stability optimisation

Stability is a good indicator of the life span of a solar cell. Multiple parameters can affect the stability of OSCs. Authors in [109] optimized these parameters using sequential minimal optimization regression on a data set obtained from the website of Danish Technical University (DTU) [110]. Authors have presented shortlisted layer-wise materials with the highest weights in sequential minimal optimization regression. These materials are the most influential materials governing the stability and performance of OPV devices [111].

Copper Content Optimization in CdTe Solar Cells

Cu is essential in CdTe solar cells as back contact and doping agent. Diffusion depth optimization of Cu resulted from diffusion annealing, and cool-down in the fabrication of CdTe solar cell was reported in [112]. ANN predicts data generated from software simulation using the Keras library in python. ANN was fed with temperature and duration of diffusing process time. Results show that the predicted and actual depths are only $0.009\mu m$ apart.

Optimization of Diode Model for Solar Cell Simulations

A bio-inspired Modified spotted hyena optimization algorithm was implemented in [113], to compare one diode model, two diode modes, and three diode model solar cells in MATLAB. The authors obtained I-V and P-V curves. They found that the three-diode model is the most accurate model.

Optimisation of Spray Plasma Processing

Optimization is a common theme in materials research when synthesizing a particular material or determining the ideal processing conditions to obtain the desired attribute. The difficulties emerge from the fact that there are several parameters whose weights might influence the outcomes. Additionally, gathering experimental data takes time and money. Authors in

[114], presented the work of [115], where BO was used to optimize the rapid plasma process. The authors used six different parameters are input that affect PCE: linear speed of pray, substrate temperature, the flow rate of precursor, gas flow rate into plasma nozzle, the height of plasma nozzle, and plasma duty cycle, while some other parameters were kept constant such as precursor formulation, concentration, etc. The optimization result showed that PCE increased from 15% to 17 %.

- ***ML for the Efficient Fabrication of Solar Cells***

Most research articles cover various ML algorithms used to fabricate PSCs effectively. However, in this section, an emphasis is given on RQ4, which examines the most optimal ML algorithms that have proven effective in identifying efficient techniques for fabricating PSCs.

PSCs are cheap to fabricate and as a result, most researchers fabricate these low-cost solar cells by trial and error. Also, fabricating a solar cell consists of a large percentage of permutations and combinations of various physical parameters such as materials used, doping layers, the thickness of the different layers, meshing, contacts, bulkiness, etc. In addition, solutions-based techniques for fabricating solar cells require less time to manufacture. However, they exhibit stability concerns. Therefore, a critical SR using the ML methods for designing a reconfigurable PSC is evaluated.

Zhe *et al.* [116] demonstrated a sequential learning architecture for producing PSCs that are guided by ML. They applied different methods to create open-air perovskite devices using the rapid spray plasma processing (RSPP) method. Further, showed the best outcome from a device made by RSPP was an efficiency improvement of 18.5% with a limited experimental budget of screening 100 process scenarios. They achieved this mainly due the three innovations such as flexible knowledge transfer between experimental processes by using prior experimental data as a probabilistic constraint, incorporation of both subjective human observations and ML insights when choosing the next experiments, and adaptive strategy of locating the region of interest using BO before conducting local exploration for high-efficiency devices.

Another research article presented by Vincent *et al.* [117] discussed a quick and simple tool for identifying the primary losses in PSCs. To comprehend the light intensity dependency of the open-circuit voltage and how it relates to the main recombination mechanism, their model used large-scale drift-diffusion simulations. The ML algorithm was developed using more than 2 million simulations and resulted in a prediction accuracy of up to 82%.

In addition, Xabier *et al.* [118] in their study used big data for the discovery of OSCs, such as non-fullerene acceptors and low-bandgap donors-based polymers. Also, they discussed the computational techniques used to choose the most promising chemical molecules from

the online material libraries. Secondly, their work provided an overview of the primary high-throughput experimental screening and characterization methodologies applicable to OSCs, specifically those based on lateral parametric gradients (measuring-intensive) and automated device prototyping (fabrication-intensive). In both scenarios, unequalled rates for the generation of experimental datasets have been achieved that leading to enhancing big data preparedness. Lastly, they used ML algorithms to locate a lucrative application to retrieve quantitative structure-activity connections and extract molecular design reasoning, which is projected to maintain the rate of material discovery in OPV.

Aaron *et al.* [119] proposed the design of experiments (DOE) and ML techniques optimizing all-small-molecule OPV cells depending on small-molecule donor, DRCN5T, and non-fullerene acceptors, ITIC, IT-M, and IT-4F. The combination was quick, efficient, and valuable resources enabled sparse but mathematically intentional reasonable sampling of huge parameter spaces. The bulk heterojunction, which is the OPV device's core layer, was optimized in this work. The optimal values of the experimental processing parameters with regard to PCE were then determined using the maps of the PCE landscape that were derived using the ML-based approach for the first and second rounds of optimization. Cagla *et al.* [120] discussed the effects of cell manufacturing materials, deposition techniques, and storage conditions on PV cell stability using a dataset containing long-term stability data for 404 organolead halide PSCs. The dataset was created from 181 published papers and analyzed using association rule mining and decision trees-based ML techniques.

Nahdia *et al.* [121] proposed an efficient method for analyzing device and material performance incorporating experimental, device modeling, and ML algorithms. The ability to implement manufacturing conditions to device performance by providing a set of electrical device characteristics results in an enlarged and faster improvement of solar energy harvesting devices. Thus, they considered parameters such as annealing temperature, surfactant selection, and charge carrier dynamics in OSCs. Followed by, Bart *et al.* [122] presented the predictions related to the bandgap of Organic Crystal Structures with the help of ML techniques. They extracted a consistent dataset of 12,500 crystal structures and the related DFT band gap properties were freely downloaded from a website. The two cutting-edge models combined yield a mean absolute error (MAE) of 0.388 eV, or 13% of the average band gap of 3.05 eV, for the ensemble. The band gap for 2,60,092 materials in the Crystallography Open Database (COD) is predicted using the trained models.

Fan *et al.* [123] presented the ML-assisted designing and fabrication of solar cells. The elements can be divided into four sub-categories: Data measurement, material properties, optimization of device architectures, and optimization of fabrication processes. The typical

types of ML techniques discussed involve ANN, GA, PSO, SA, RF, etc. Among them, ANN and GA are the two ML techniques that are most frequently used.

2.5 Summary

In summary, a critical literature review is presented regarding the prediction of solar irradiance using ML techniques and the ML framework of the LCOE and the ROI of a PV system. The presented literature review tends to cover the summary of case studies of the research articles that used various ML techniques to forecast the LCOE and ROI parameters. Moreover, the literature review summarises that previous studies attempted to use singular inputs for calculating the LCOE and ROI of a PV system which resulted in poor estimation of the key parameters of the PV system. This information is vital in suggesting the gap to accurately predict the LCOE and ROI of a PV system to indicate these parameters to the client.

This chapter also evaluated a broad range of ML techniques for optimizing the performance of low-cost solar cells. The SR in this chapter indicates that a significant proportion of research focuses on data-driven approaches and ML techniques for discovering low-cost solar cells, with a third of publications targeting ML algorithms in the fabrication process. The SR suggests that ML techniques can potentially accelerate the discovery of new solar materials and architectures. Future research can expand on these findings by exploring and developing new ML techniques for solar cell optimization. Additionally, addressing the scalability and sustainability of low-cost solar cell technologies to enable large-scale commercialization is essential. Ultimately, applying ML techniques in solar energy can revolutionize the industry and pave the way for a cleaner and more sustainable future.

It is worth mentioning here that the SR presented in the thesis reviews the research papers that use ML techniques to discover low-cost, high-performance solar cells. Moreover, the review results presented in the SR are of significant importance to the researchers in building an ML model capable of predicting an efficient solar cell architecture. Accordingly, in this thesis, the SR is presented; however, the following next two chapters predominantly discuss the ML models that are capable of predicting solar irradiance and the provide an accurate method of estimation of LCOE and ROI parameters of a PV system.

Chapter 3

Machine Learning for Predicting Solar Irradiance

3.1 Introduction

Over the decade, there has been rapid growth in the field of the Internet of Things (IoT) due to its improved connectivity, data-driven decision-making, and enhanced customer experiences. The IoT has transformed the lives of people to a large extent by efficiently connecting people across the world and turning the planet into a much smarter and more advanced globe [124]. In addition, IoT has the potential to profoundly enhance wireless networking technologies due to its increased bandwidth, connectivity, improved network coverage, increases use of network resources, enhanced security, and greater scalability.

One of the subsets of IoT is the Wireless Sensor Network (WSN) which uses a combination of sensors to wirelessly interact and communicate with other sensor nodes. Moreover, the architecture of a WSN comprises a gateway node (central) connected with multiple sensor nodes (branches) to share information in the form of data packets from the transmitter to the receiving end [125]. The WSN with the help of sensor nodes, senses, gathers, processes, and transmits information such as weather sensing, the healthcare industry, smart grids and robotics.

The energy consumption in a WSN is due to sensing energy (sensors), computing energy (data processing) and communication energy (a short radio-frequency circuit that performs data transmission and reception) [126]. The WSNs rely on batteries to feed power to the sensor nodes and gateway. However, battery capacity is limited due to a mismatch between supply and demand. In addition, WSN nodes are installed in remote locations such as deserts, forests, war zones and in seas, [127] where human access is often restricted or limited.

Moreover, frequent replacement of the battery is not possible in these remote locations [128] and thus, the energy management of the WSN nodes plays a vital role in maintaining the prolonged operation with minimal investments. Therefore, the proposed model aims to develop solar energy harvesters that scavenge renewable energy from the surroundings and predict the solar irradiance ahead of time [129].

Another limitation of energy harvesting is the intermittent nature of renewable resources. Though the sensor nodes are expected to consume less amount of energy, however, the main concern is the energy fluctuation and variable DC output from the solar cells. Not only does the energy vary in sensor nodes but also it varies within the renewable energy resources due to their intermittent nature [130]. For instance, solar energy is available during the day only and wind energy varies with the wind speed. Hence, the chapter proposes a model aiming to harvest renewable energy adequately from the environment overcoming the variations in renewable energy availability. Thus, predicts the amount of renewable energy using machine learning (ML) algorithms.

ML is a field of research that allows the machine to learn using past experiences and thus, train the machine to predict the future or possible outcomes. The ML comprises three types, supervised learning, unsupervised learning, and reinforcement learning. Also, the model uses supervised ML to predict the amount of solar irradiance generated on the WSN using historical data. Further, divided the data to train and test the model using multiple solar irradiance parameters such as global horizontal irradiance, direct normal irradiance, ambient temperature, humidity, and latent heat of flux of the chosen location for ten years.

Also, performed feature extraction using correlation analysis of several parameters of solar irradiance such as global horizontal irradiance, ambient temperature, humidity, latent heat of flux, and normal direct irradiance. Further, analysis of the model is done with the help of feature scoring of the multiple features representing their dependency on each other using correlation analysis and heat map. In addition, to estimate the hyper-parameters in multiple output support vector regression (SVR), the model performed the grid search to find the optimized values of each hyper-parameter, i.e., (C, Gamma, and Kernel). Moreover, the results represent a reduced computational complexity for determining the hyper-parameters for varied time series. In the last section, a comparison of the results from Linear Regression, Single Output SVR, and Multiple-Output SVR in terms of R2-Squared value, Root Mean Square Error (RMSE), and Mean Absolute Error (MAE) performances are conducted.

From this chapter, an attempt has been made to accurately predict the amount of solar irradiance using the WSN installed at remote locations and forecast the availability of solar irradiance using two ML algorithms i.e., linear regression model and the SVR model.

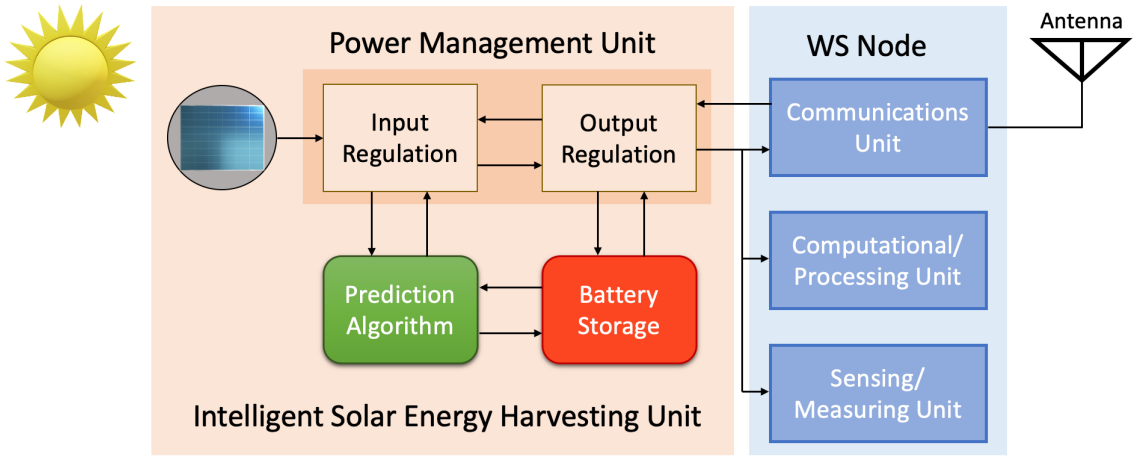

Fig. 3.1 The block diagram of the solar powered WSN and its constituent components.

3.2 System Model

Figure 3.1 represents the block diagram of the WSN and its constituent components.

The solar-powered WSN scavenges energy from the surrounding, here from the sun. Then, the amount of solar energy generated (in the form of data packets) communicates with input regulation (indicating the WSN about the input solar energy), battery storage, prediction algorithm and output regulation (indicating the amount of output energy after charging the battery) to intelligently predict the amount of solar irradiance in the Power Management Unit of the Intelligent Solar Energy Harvesting Unit. Further, the output from the power management unit is fed to the connected WSN node and antenna to share information with other sensor nodes. In addition, the next sections of the chapter discuss the prediction algorithms for accurately forecasting solar irradiance using ML techniques.

Figure 3.2 represents the working model of the proposed WSN architecture using ML techniques. The vector $\mathbf{X} = \{X_1, X_2, ..., X_n\}$ is the input to the ML algorithm, which is used to train and test the model to predict the solar irradiance, $\mathbf{h} = \{h_1, h_2, ..., h_n\}$ is weight vector of the SVR hidden layers, and Y is the final forecasted solar irradiance that is the input to the main node of the WSN.

3.2.1 Data Processing

For the study, the dataset is extracted for two locations, Sacramento, California, USA, and Delhi, India having global coordinates as (38.58, -121.35) and (28.58, 77.16,) respectively. The dataset for 2010 to 2020 is collected and divided so that the analysis is performed on the time interval of 1 week, 1 month, 1 year, and 10 years having an hourly time resolution. The

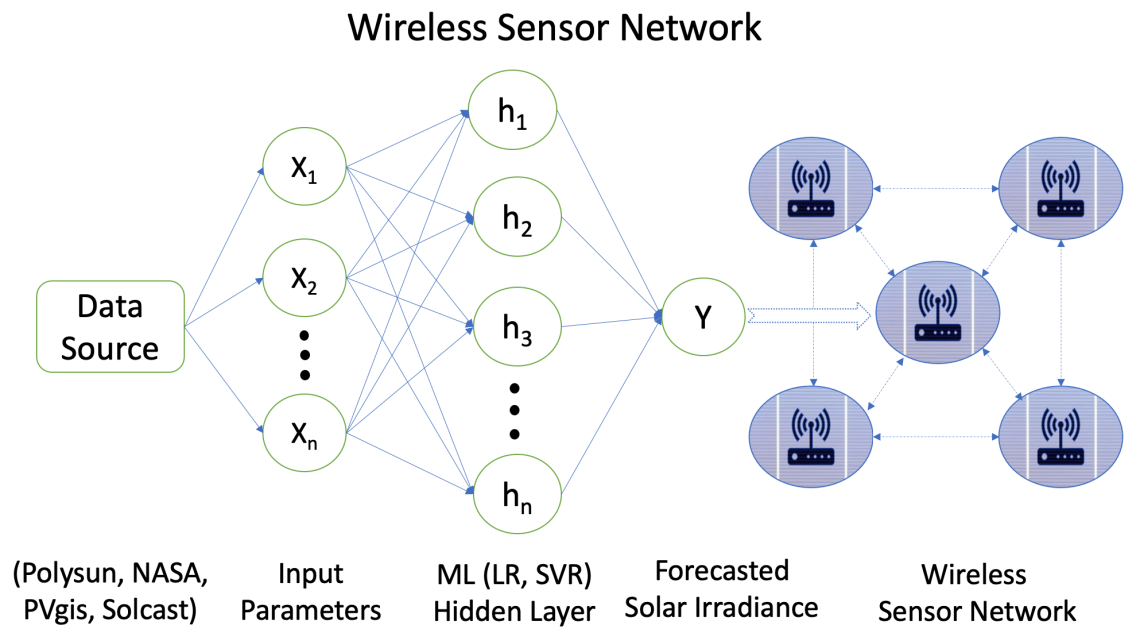

Fig. 3.2 Working Model of the proposed WSN architecture using ML techniques.

dataset is accessed from the licensed version of the software Polysun [131] which provides solar irradiance values using the appropriate sensors installed at some particular locations in hourly resolution for a period of 1 year (randomly chosen year between 1996 to 2015).

Figure 3.3, depicts the information on statistical as well as historical graphs of the dataset and scattered data plot depicting the correlation of parameters to each other. The dataset consists of parameters such as global horizontal irradiance (Wh/m2), ambient temperature (ºC), latent heat of flux (W/m2), humidity (%), and direct normal irradiance (Wh/m2) for the location Sacramento, California, the USA for the year 2015.

3.2.2 Identification of Outliers in the dataset

For the SVR problems, the outliers play a significant role in determining the best fit line or hyperplane, thus it is important to identify the outliers in the dataset for each parameter. The proposed model evaluated four input parameters, global horizontal irradiance, latent heat of flux, ambient temperature, and humidity for the location California, the USA for 1 year. Figure 3.4 represents the box-plot of outliers in the SVR for each input parameter independently.

In addition, the parameter global horizontal irradiance has the maximum number of outlier which means that there is alot of noise in the dataset. Hence, it is useful to neglect the values of global horizontal irradiance which are greater than 300 W/m2. Moreover, there is

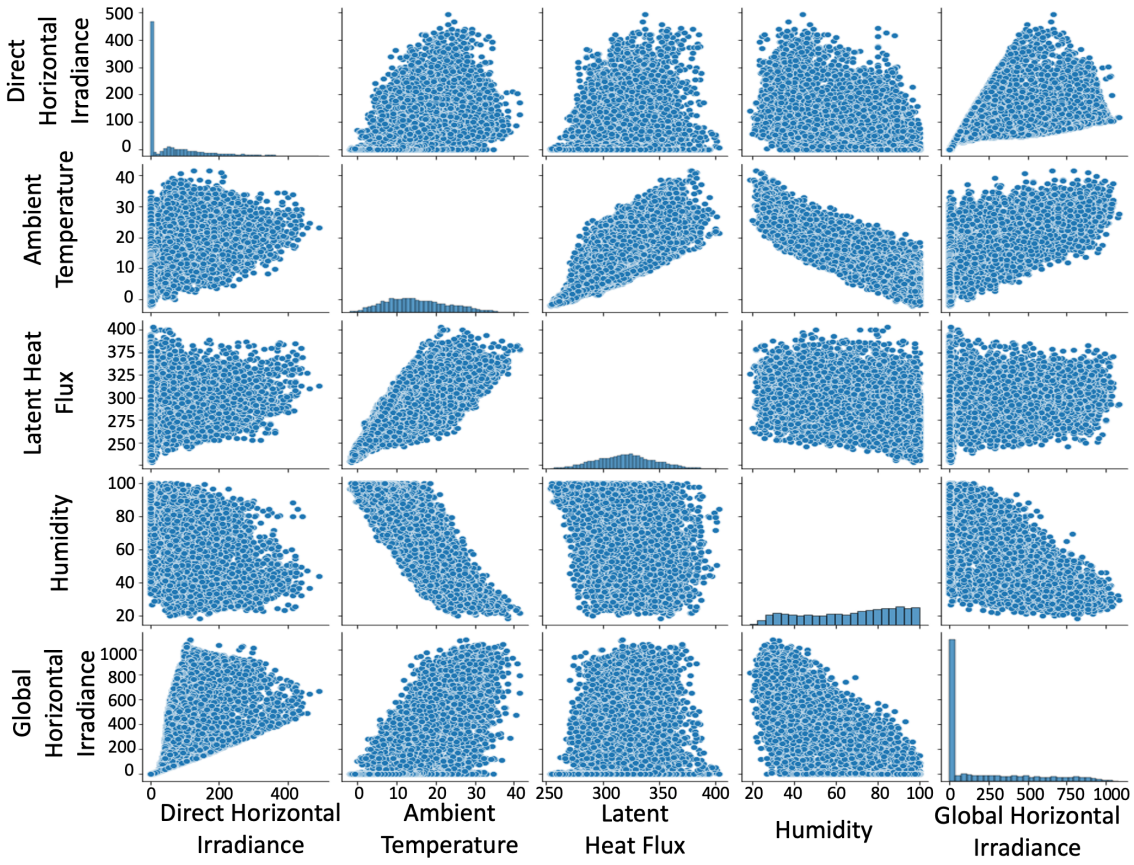

Fig. 3.3 Statistical representation for the solar irradiance dataset for California, USA.

some noise in the parameter latent heat of flux and ambient temperature, however, there are significantly lower values and thus, will not affect variance in the dataset.

3.2.3 Feature Engineering

Further, the feature engineering is performed to determine the importance of each parameter with respect to each other. The feature selection consists of correlation analysis and principle component analysis (PCA). For the study analysis, the correlation analysis using the heat map is performed as shown in the figure 3.5. The direct normal irradiance has the maximum percentage of importance on output results, followed by ambient temperature and latent heat of flux. Further, the humidity has the least importance on the output results, hence, in the ML algorithm does not considers humidity as an input variable.

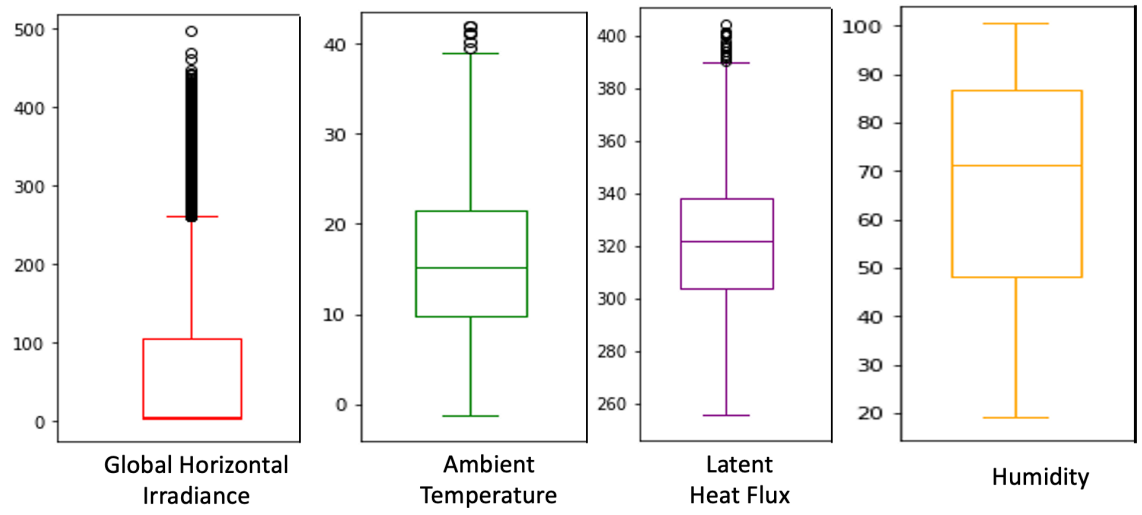

Fig. 3.4 Identification of outliers in the dataset for the parameters global horizontal irradiance, latent heat of flux, ambient temperature and humidity.

3.3 Model Training

3.3.1 Linear Regression Model

Mathematically, the linear regression model is a statistical approach of predicting the value of the cost function depending upon another variable. The linear regression model uses past experiences of the parameter as an input and predicts the outcome having a linear relationship between the cost function and output variables [132]. The linear regression model is implemented because of the inter-dependency of multiple features on the prediction of solar irradiance.

For the proposed model, linear regression with multiple features is used for forecasting solar irradiance. The 4 parameters were processed from the dataset of the location California, the USA for 1 year and further divided into training (80%) and testing (20%) for the machine to learn in Python.

3.3.2 Suppport Vector Regression Model

The Support Vector Machine (SVM) is an ML algorithm that consists of two types of paradigms, classification and regression problems. After analysing the results from linear regression, an attempt has been made to implement the SVR algorithm to predict global solar irradiation for making critical comparisons between varied ML techniques. Like linear regression, the SVR is also classified as a Supervised ML algorithm that is used to determine

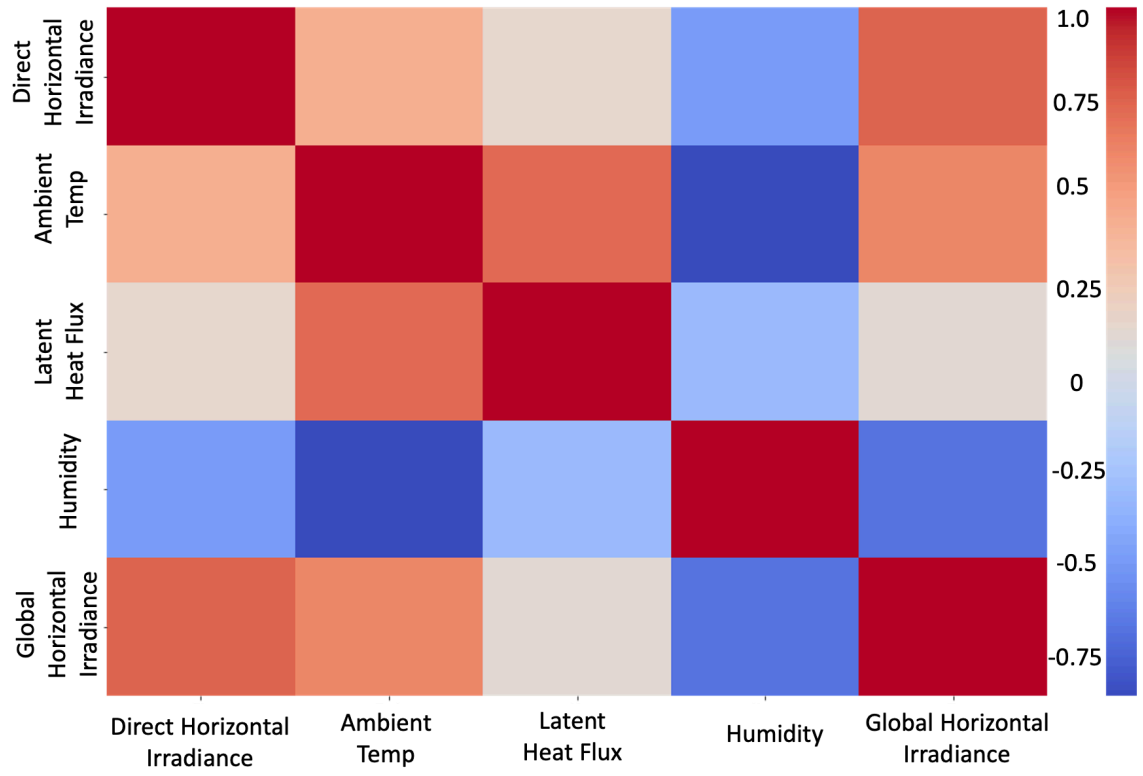

Fig. 3.5 Heat Map determining the correlation of the feature with respect to each other.

the best possible fit line (linear) or a hyperplane (non-linear) containing the maximum points in a dataset [133].

Further, the optimized decision boundary for a line or a hyperplane can be calculated considering the values of the most appropriate hyperparameters. In general, there are three hyper-parameters C, kernel, and gamma. First, the C, also known as the punishment factor of the error term, helps to evaluate the significance of outliers in the dataset. Second, the kernel is decided based on the dimensions of the dataset, for instance, if the decision boundary under consideration is linear or non-linear (radial basis function is used for multi-dimension kernel). Third, the gamma hyper-parameter for a non-linear function (uses radial basis function) influences the distance from a single train point in the dataset.

Similar to the linear regression model, the dataset is divided into the train (80%) and test (20%) and implemented the SVR model to forecast global solar irradiance. However, SVR is also used for estimating the values of C, gamma, and kernel. Accordingly, the values of C are assumed initially to be in the range of 0.01, 0.1, 1, and 10. On the contrary, the values of gamma were assumed to be 0.001, 0.01, 0.1, and 1 as well as for the kernel; it was linear or rbf. The SVR uses the Grid Search function to determine the most optimized values for C, gamma, and kernel.

3.4 Results and Discussion

3.4.1 Forecasting Solar Irradiance using LR Model

Figure 3.6 depicts the plot of predicted global solar irradiance (orange curve) to the actual values (blue curve) of the solar irradiance for the location CA, USA. The overall accuracy of the model was 81.92%. In addition, linear regression with multiple features was also incorporated for another location i.e. Delhi, India, and the overall accuracy was 87.30%.

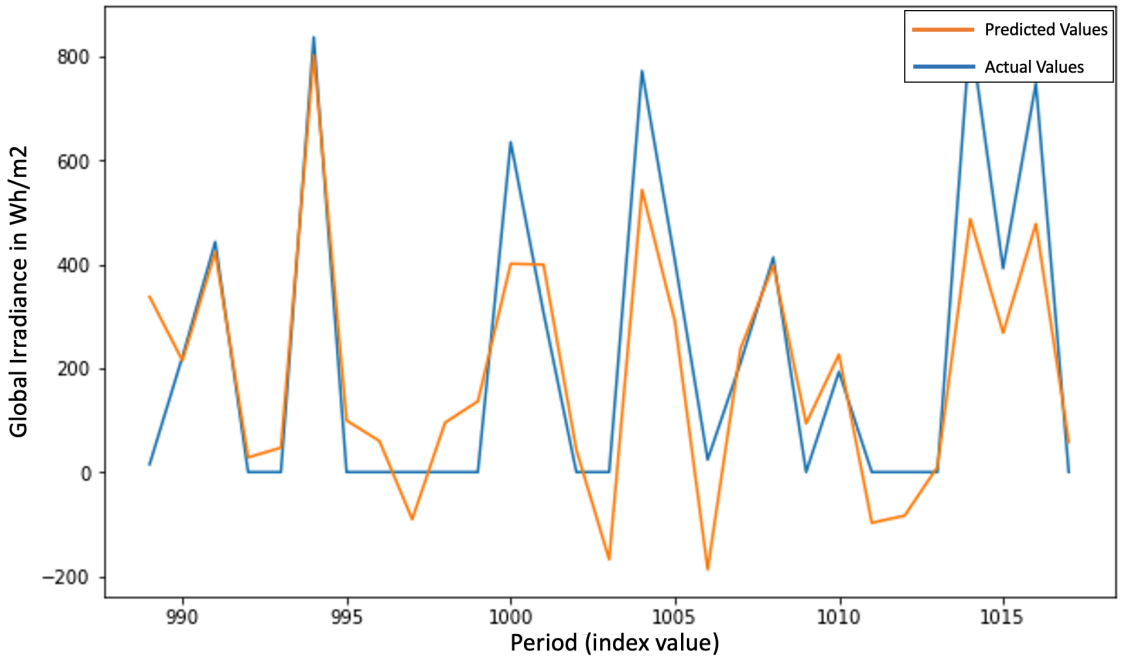

Fig. 3.6 Predicted solar irradiance using Linear regression for multiple parameters and zoom image of actual vs predicted curve.

3.4.2 Forecasting Solar Irradiance using SVR Model

Further, implementing the grid search for the calculations of optimized values of hyper-parameters, the results obtained were, C = 10, Gamma = 0.001 and kernel = rbf. Further, incorporating these hyper-parameters as input to the SVR model, the dataset was divided into train and test of 80% and 20% respectively.

Likewise, figure 3.7 represents the actual values vs forecasted solar irradiance using the SVR model. Here, the model under consideration observed that training the dataset precisely, following the trend, and thus, forecasted the solar irradiance with an accuracy of 81.6% (also

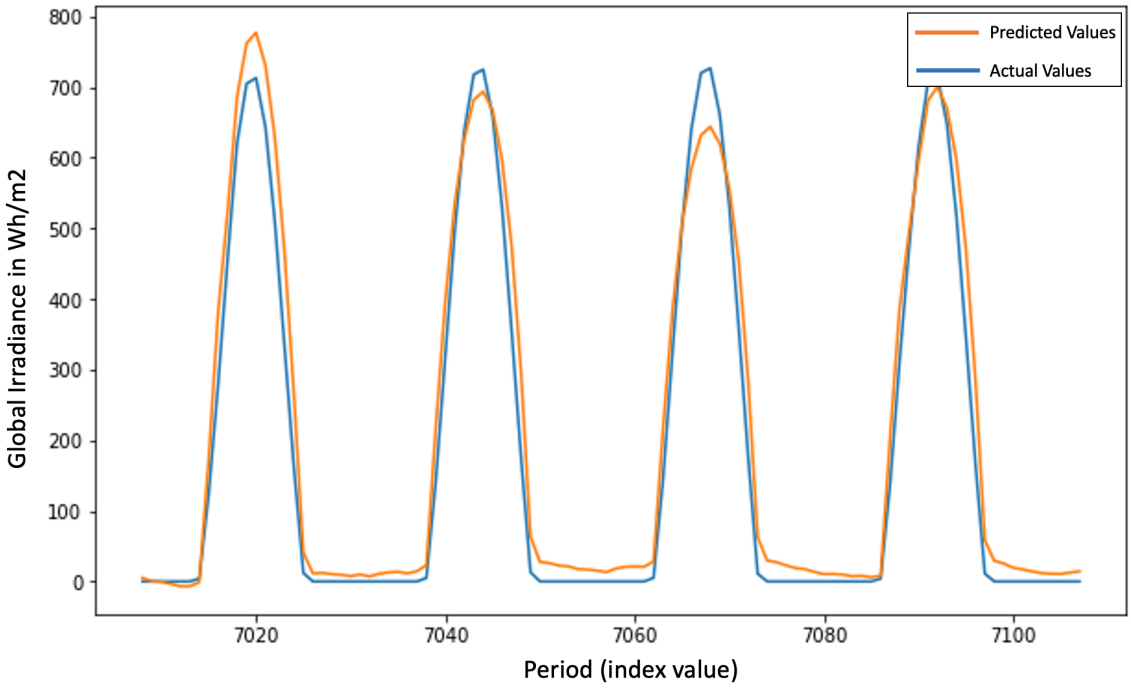

Fig. 3.7 Predicted solar irradiance using Multiple Parameter SVR and zoom image of actual vs predicted curve

known as the R-Square Performance value). The RMSE and MAE of the model are 6.9% and 3.1% respectively.

Table 3.1 Comparison of results based on RMSE, R-Squared value and the MAE for a dataset of 10 Years.

Hyperparameter	**California, USA**	**Delhi, India**
"C"	10	10
"Gamma"	0.001	0.001
"Kernel"	rbf	rbf
"R-Squared (%)"	39.68	40.21
"RMSE (%)"	21.7	24.12
"MAE (%)"	13.07	17.9

Table 3.2 Comparison of results based on RMSE, R-Squared value and the MAE for a dataset of 1 Year.

Hyperparameter	**California, USA**	**Delhi, India**
"C"	10	10
"Gamma"	0.01	0.01
"Kernel"	rbf	linear
"R-Squared (%)"	81.6	82.11
"RMSE (%)"	6.9	8.2
"MAE (%)"	3.1	4.6

Followed tables, 2.1, 2.2, 2.3, and 2.3, represent the optimized values of hyper-parameters (C, gamma, and kernel) calculated using the grid search for the data sets of 10 years, 1 year, 1 month and 1 week, respectively. Also, the tables include the R-squared value, RMSE, and MAE for the locations Sacramento, California, USA, and Delhi, India, for 10 years, 1 year, 1 month, and 1 week. According to the achieved results, the SVR model is expected to perform well if the R-square value is above 80% and RMSE, as well as the MAE value of the model, is less than 10% and 5%, respectively. Therefore, the Multiple output SVR for 1 year and 1 month of California, USA, and Delhi, India, has much better forecasting results as compared to the duration of 10 years and 1 week.

3.5 Summary

In summary, the chapter introduces an efficient method to power WSN devices by developing WSNs with energy harvesters that are capable to scavenge solar energy. Further, the amount

Table 3.3 Comparison of results based on RMSE, R-Squared value and the MAE for a dataset of 1 Month.

Hyperparameter	**California, USA**	**Delhi, India**
"C"	10	10
"Gamma"	0.001	0.001
"Kernel"	rbf	linear
"R-Squared (%)"	76.47	84.13
"RMSE (%)"	5.6	9.2
"MAE (%)"	2.4	6.8

Table 3.4 Comparison of results based on RMSE, R-Squared value and the MAE for a dataset of 1 Week.

Hyperparameter	**California, USA**	**Delhi, India**
"C"	1	10
"Gamma"	0.01	0.001
"Kernel"	linear	linear
"R-Squared (%)"	82.59	72.16
"RMSE (%)"	14.9	15.95
"MAE (%)"	7.9	8.3

of solar irradiance is predicted using the ML algorithms, so that the node can intelligently manage its energy independently. The chapter analyzes data by executing data processing, identifying outliers, and feature engineering to determine the most significant values of solar irradiances and respective parameters for two locations, i.e. California, USA, and Delhi, India. Also, a grid search approach was performed to find an optimized value of hyper-parameters, such as C=10, gamma = 0.001, and kernel = rbf for California, USA. Therefore, performing these optimization techniques helps to improve the overall prediction performance for estimating the amount of solar irradiance. Moreover, the ML model showed that the output results from the SVR model predicted the solar irradiance more accurately as compared to the linear regression model. In addition, the compelling results were that the models with a time duration of 1 year and 1 month have much better forecasting results than 10 years and 1 week, with both RMSE and MAE less than 7% for Sacramento, California, USA.

Chapter 4

Machine learning framework for LCOE of PV System

4.1 Introduction

World's energy demand is growing fourfold and accordingly, most of the world's energy is consumed by China and the United States of America and is followed by India. [134, 135] Subsequently, these escalating energy needs can be achieved using solar energy, which can potentially improve the lives of communities profoundly worldwide [136, 137]. Harnessing the Sun's renewable, sustainable and low-carbon energy source can be achieved using PV systems [138]. The PV systems convert sunlight directly into electricity and can often do so with high efficiency [139]. Among the challenges in achieving widespread adoption of this technology is the price of solar electricity in comparison to conventional sources of energy [140].

Grid-connected PV systems are cost-effective renewable energy solutions that do not require batteries and mainly consist of a PV array generator and an inverter for converting DC electricity to AC [141, 14]. Furthermore, they can be configured to supply energy to primary loads, with all excess electricity being sold to the grid or even bought back from the grid when the PV supply is insufficient [142]. Alternatively, all the energy generated from the PV system can be sold directly to the grid [143]. In all cases, the difference between the price of buying electricity from the grid and that of selling to the grid from the PV system is a substantial factor in determining the optimal size of grid-connected PV systems [144].

Therefore, evaluating the economic feasibility of a PV system is extremely important [145]. For example, users need to know their expected return on investment (ROI), and funding agents need means to analyze proposed technology development [146]. Similarly,

technology developers need to understand how they will compete relative to other technologies [147]. Moreover, regulators and policymakers (who help define the economics of energy production) require reliable information [148]. The capital cost of a PV system, its operation and maintenance costs and its expected energy yield must be considered systematically so that a comparison with conventional fossil fuels can be made [149]. Consequently, one needs a method to compare energy costs fairly. Therefore, a generalised framework is developed for predicting the feasibility of a grid-connected PV system for utility-scale applications. The prediction framework considers the amount of electricity consumption using several metrics, including Gross Domestic Product (GDP), prices of electricity, population growth and weather data.

The chapter of the thesis is divided into 6 sections. Section 4.2 reviews the literature on calculating the LCOE of grid-connected PV systems. Section 4.3 describes the methodology for calculating the LCOE and ROI of a grid-connected PV system using various ML algorithms and briefly discusses the proposed model. Section 4.4 includes the data explanation and steps of data pre-processing. Section 4.5 presents the results of the energy prediction models using various ML techniques and provides comparisons with singular input demographic variables. Next, section 4.6 includes a discussion of the results, and lastly, section 4.7 concludes the results from the various ML algorithms that are applied throughout the chapter.

4.2 Methodology

The methodology section of the chapter demonstrates all the necessary calculations for calculating the LCOE and ROI. Moreover, the section includes information on the ML algorithms used for forecasting the LCOE using demographic variables. In addition, an ML model is applied for estimating the LCOE and, therefore, the ROI of the utility-based grid-connected solar home system installed in Sacramento, California, USA and the Philippines.

4.2.1 Calculating the LCOE

A valuable parameter for comparing the cost of electricity production from any energy generation system over its lifetime is the Levelized Cost of Electricity (LCOE) [150]. This is typically defined as the average cost ($) per kWh of useful electrical energy generated by the system throughout its years of operation. Mathematically, the LCOE can be calculated as follows:

$$\textbf{LCOE} = \frac{\text{System Lifetime Cost, } L_t}{\text{Lifetime Energy Production Cost, } E_t} \tag{4.1}$$

$$L_t = I_t + C_t + S_t \tag{4.2}$$

$$E_t = E_o(1-d)^t \tag{4.3}$$

where, I_t represents the initial costs, including expenses related to equipment, land and other setup necessities. The total costs paid at the beginning of the project, such as annual operation and maintenance costs, are denoted as C_t. Lastly, S_t signifies the salvage value, which is the use value of the project at the end of its lifetime. Similarly, the Lifetime Energy Production Cost (E_t) is defined as $E_t = E_0(1-d)^t$, where E_0 refers to the initial rated energy output and the system degradation factor is represented by $(1-d)^t$. This formula accounts for the gradual decrease in energy output over time due to factors such as aging and wear of the PV system components.

Traditional methods of calculating LCOE relied on using singular input values for each of the variables above [151]. For instance, using the benchmark prices reported for 2017, a 50 MW utility-scale PV power plant installed in California would cost $56 million, corresponding to $1.12/W (31% module, 69% balance of systems). This system would produce approximately 86 GWh of energy in the first year. Assuming that the discount rate is 5.5%, the federal tax rate is 30%, the state tax rate is 8%, the evaluation period is 25 years, and the system degradation is 0.5%, then the LCOE of this system is 5.83 c/kWh. A careful consideration of these numbers, as mentioned above, shows that many assumptions have already been made to determine the LCOE of this system.

Nevertheless, a case study of a PV system installed in Spain indicates that estimating the LCOE using traditional methods may lead to inaccurate estimations. This is particularly relevant as factors such as inflation rate, discount rate, degradation rate, and Consumer Price of Electricity (CPE) are likely to vary during the lifetime of a PV project (typically 25 years). In Spain, LCOE analysis proved inadequate when an excessive number of projects were developed based on overly optimistic assumptions regarding panel failure rates and other performance factors [152]. A more comprehensive examination of the uncertainties associated with these assumptions might have averted significant losses.

4.2.2 Calculating the ROI

After determining the LCOE, the next step is to calculate the ROI for the PV system. To do this, it is necessary to compare the LCOE with conventional electricity prices [153]. However, this parameter is likely to change over the project's lifetime. As the input parameters continuously change and are strongly dependent on the system's location, accordingly, ML

techniques are proposed to determine the ROI of a PV system accurately. ML techniques can effectively capture the complex relationships between various factors and adapt to changing input conditions, making them well-suited for predicting the ROI of the PV system with higher accuracy.

Mathematically, the ROI of a PV system can be calculated using,

$$ROI = T_C / B_I \tag{4.4}$$

where T_C represents the *Total Cost of the PV System*, and B_I denotes the annual benefit from the installation of the PV system. Here, the Total Cost of the PV system refers to the initial investment required for the PV system, including costs related to equipment, land, installation and other setup necessities. It is also sometimes called the Capital Expenditure (CAPEX) cost. Therefore, the ROI parameters estimate the number of years a client can expect to achieve a return on investment for installing a PV system. A review of the literature reveals that the methods for calculating the ROI used by researchers worldwide often rely heavily on assumptions, leading to imprecise cost analysis estimations [154].

4.2.3 Machine Learning Implementation

Due to the above-mentioned factors, several ML techniques for calculating the LCOE and ROI of the PV system are examined. Subsequently, in this subsection, the ML techniques discussed can be applied to forecast the LCOE accurately. ML is a branch of computer science that involves computational training algorithms to make predictions based on a known input data set [155]. Furthermore, ML can be divided into three main categories named Supervised ML, Unsupervised ML, and reinforcement ML algorithms [25]. In supervised ML, the computer learns from the input provided by the user, whereas in unsupervised ML, the computer learns patterns from untagged data [156]. Moreover, reinforcement ML is the technique of the computer to learn from the hit-trial-error method [20]. For the study under consideration, the proposed model mainly relies on Supervised ML using regression techniques to predict the LCOE of PV systems, followed by the calculations for the ROI of the PV system shown. Lastly, the LCOE calculated by different approaches is compared using fixed input values for the various parameters.

4.2.4 Long Short-Term Memory (LSTM) Model

Long Short-Term Memory, often known as LSTM, is a sort of recurrent neural network (RNN) that was created expressly to deal with long-term dependencies and prevent the

vanishing gradient problem that can happen in conventional RNNs. A popular deep learning architecture called an LSTM model is employed in many sequence-based prediction applications, including natural language processing (NLP), speech recognition, and others. An LSTM model, at its most basic level, is made up of a series of LSTM cells that are linked to one another in a chain-like fashion. Three major parts make up an LSTM cell: a memory cell, an input gate, and an output gate. The input and output gates, respectively, control the flow of information into and out of the cell, while the memory cell is in charge of storing data about the prior inputs [157].

The output gate determines what data should be output from the memory cell, and the input gate determines what data should be admitted into the memory cell. Both gates are managed by activation functions, which are masterable by training. Additionally, each memory cell includes a set of activation functions that control how it updates its state in response to input from the present and information from the past. The input sequence is fed into an LSTM model's forward pass one element at a time, and each cell's output is given to the cell after it in the sequence. A prediction about the sequence as a whole is then made using the output of the last cell. The LSTM model's parameters are altered during the training phase using an optimization technique like stochastic gradient descent (SGD) or Adam. For each input sequence in the training set, the objective is to reduce the discrepancy between the projected output and the actual output.

4.2.5 Autoregressive Integrated Moving Average (ARIMA) Model

A statistical model called ARIMA (AutoRegressive Integrated Moving Average) is used for time series analysis and forecasting. It is a class of models that are frequently used in finance, economics, and other disciplines where time series data is prevalent to depict a variety of time series patterns. The moving average (MA), integrated (I), and autoregressive (AR) components are the three parts that makeup ARIMA models.

The AR component simulates the relationship between the time series' historical values and the current value. In particular, the AR component makes the assumption that the time series' present value is a linear function of its earlier values, with the linear function's weights assessed during model fitting. The time series' differencing, which is employed to make it stationary, is modeled by the I component. A stationary time series is simpler to model and analyze since it has a constant mean and variation across time. To get rid of any pattern or seasonality in the data, differencing includes subtraction of the prior value from the present value. The MA component simulates how the present value depends on the historical time series mistakes.

The theory underlying this component is that past forecasting errors, which can be represented as a weighted sum of past errors, have an impact on the current value of the time series. Three parameters—p, d, and q—are generally used to define ARIMA models. The q parameter denotes the order of the MA component, the p parameter denotes the order of the AR component, and the d parameter denotes the order of differencing required to keep the time series stationary. The model parameters must be determined using a procedure known as parameter estimation in order to fit an ARIMA model to a time series. This entails determining the values of p, d, and q that best match the time series data using a statistical technique like maximum likelihood estimation or least squares. The model can be employed to predict future values of the time series once the parameters have been evaluated [158].

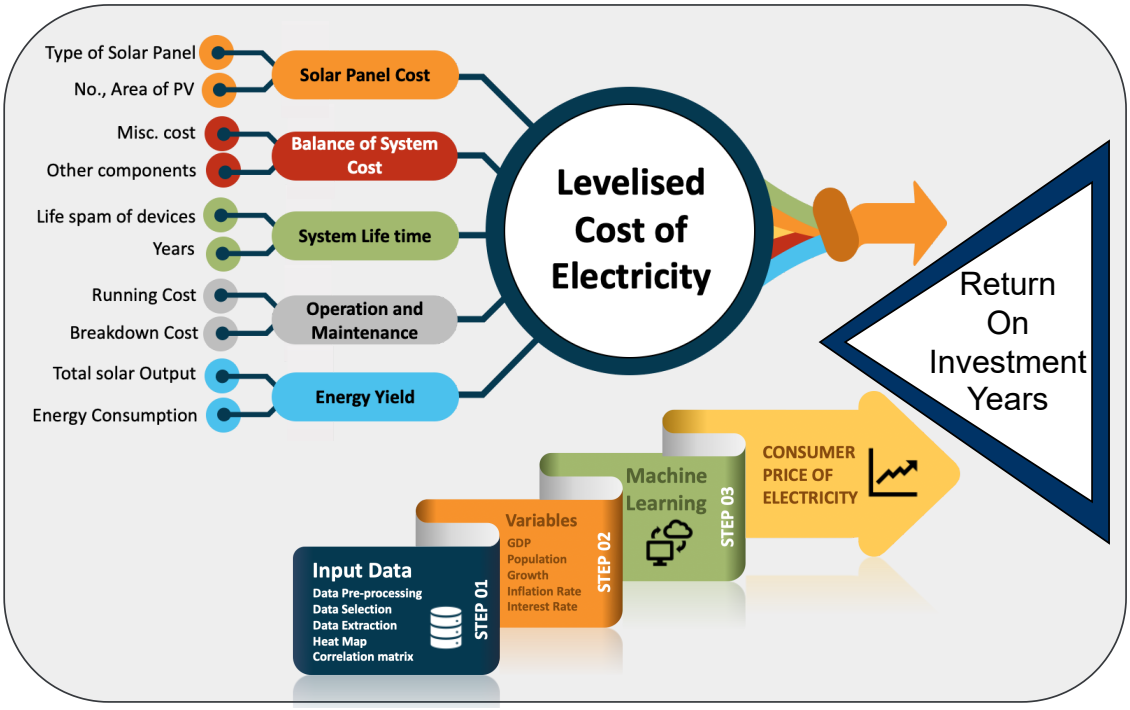

Fig. 4.1 The proposed model for determining the LCOE and ROI for a utility-connected solar home system.

4.2.6 Proposed Model

In a majority of previous studies, researchers have calculated the LCOE and ROI using singular input values, typically assuming that the CPE will increase by 5-10% over the lifetime of the solar plant. On the contrary, the chapter argues that the estimation of LCOE should account for variations in CPE due to factors such as population growth, inflation rate and interest rate over time [159]. Consequently, an algorithm is proposed that accurately considers

these dynamic variables. Using historical data, various ML algorithms are incorporated to estimate the LCOE, considering the aforementioned factors. To validate the proposed model, the historical data is extracted from two regions, California in the USA and Butuan City in the Philippines and compute the error function of various ML techniques to identify the most suitable ML model. Figure 4.1 illustrates the proposed system, encompassing the required input parameters, relevant variables and ML algorithms used to determine the ROI of a PV plant.

4.3 Data Explanation

This section of the chapter explains detailed information about the selected dataset values for accurately estimating the LCOE and ROI of utility-based grid-connected solar home systems. The data extraction step is initially thoroughly discussed, followed by a plot of datasets and heatmap for independent and dependent variables determining the correlation matrix.

4.3.1 Data Extraction

To predict the LCOE and ROI, it is crucial to obtain real-time data from reputable sources such as the Environmental Investigation Agency (EIA) [160], International Renewable Energy Agency (IRENA) [161], Bureau of Economic Analysis (BEA) [162], and International Energy Agency (IEA) [163]. Moreover, the consumer price of electricity data was extracted from EIA, followed by the statistical data on population growth and the gross domestic product extracted from the websites BEA and IRENA. It is worth mentioning that some of the data were available on a quarterly or annual scale; however, to maintain the unity in the data comparison, the data was extrapolated using Python's Generative adversarial networks (GAN) framework to obtain the complete data on an annual scale. For the study, historical data for Sacramento, California, USA, is collected. The datasets comprise independent demographic variables such as the Consumer Price Index (CPI) as a measure of the inflation rate (X_1), population growth (X_2), and Gross Domestic Product (GDP) (X_3). In contrast, the dependent variable (Y) is represented by the average CPE (cents/kWh) [164]. The timescale for the extracted dataset is monthly. Table 1 presents the respective dependent and independent variables' dataset of demographic values, ranging from January 2005 to December 2021.

Table 4.1 The table showcases an example of data extracted from various online websites such as EIA, IRENA, BEA, IEA, etc.

Date	Inflation (X1)	Population Growth (X2)	Gross Domestic Product (X3)	CPE (Y)
01/01/2005	200.35	294957.00	128234.50	12.19
01/02/2005	201.20	295167.33	128717.47	12.33
01/03/2005	201.85	295377.67	129200.43	12.12
01/04/2005	202.50	295588.00	129683.40	12.57
01/05/2005	201.85	295838.67	130635.33	13.4
01/06/2005	201.20	296089.33	131587.27	13.16
01/07/2005	202.10	296340.00	132539.20	13.43
01/08/2005	203.00	296588.67	133226.70	12.14
01/09/2005	204.45	296837.33	133914.20	11.3
01/10/2005	205.90	297086.00	134601.70	11.28
01/11/2005	204.65	297302.67	136682.93	12.8
01/12/2005	203.40	297519.33	138764.17	12.91
. . .	. . .	. . .	. . .	. . .
. . .	. . .	. . .	. . .	. . .
01/01/2021	303.67	331949.00	312120.20	21.43
01/02/2021	304.39	331973.00	310499.27	22.53
01/03/2021	306.90	331997.00	308878.33	23.37
01/04/2021	309.42	332021.00	307257.40	22.75
01/05/2021	309.46	332113.00	309934.50	23.11
01/06/2021	309.50	332205.00	312611.60	22.46
01/07/2021	310.33	332297.00	315288.70	23.34
01/08/2021	311.17	332392.67	319178.13	23.44
01/09/2021	312.22	332488.33	323067.57	21.97
01/10/2021	313.27	332584.00	326957.00	22.77
01/11/2021	314.54	332639.00	327936.67	23.83
01/12/2021	315.81	332694.00	328916.33	23.22

4.3.2 Statistical representation of Dataset

Initially, the collected raw dataset is non-uniform and has noise, disturbances, irregularities, seasonality, trends or patterns associated with them. Therefore, it is essential to understand these parameters before inputting them into the ML model. Subsequently, Figure 4.2 depicts the plot of dependent and independent variables used to estimate the dependent variable, i.e. the CPE. The dataset plot shows that the parameters such as population growth, the consumer price of the index, the GDP and the CPE have a linear relationship. However, parameter CPE has seasonality, noise, and irregularities associated. The range of chosen dataset is from January 2005 to December 2021.

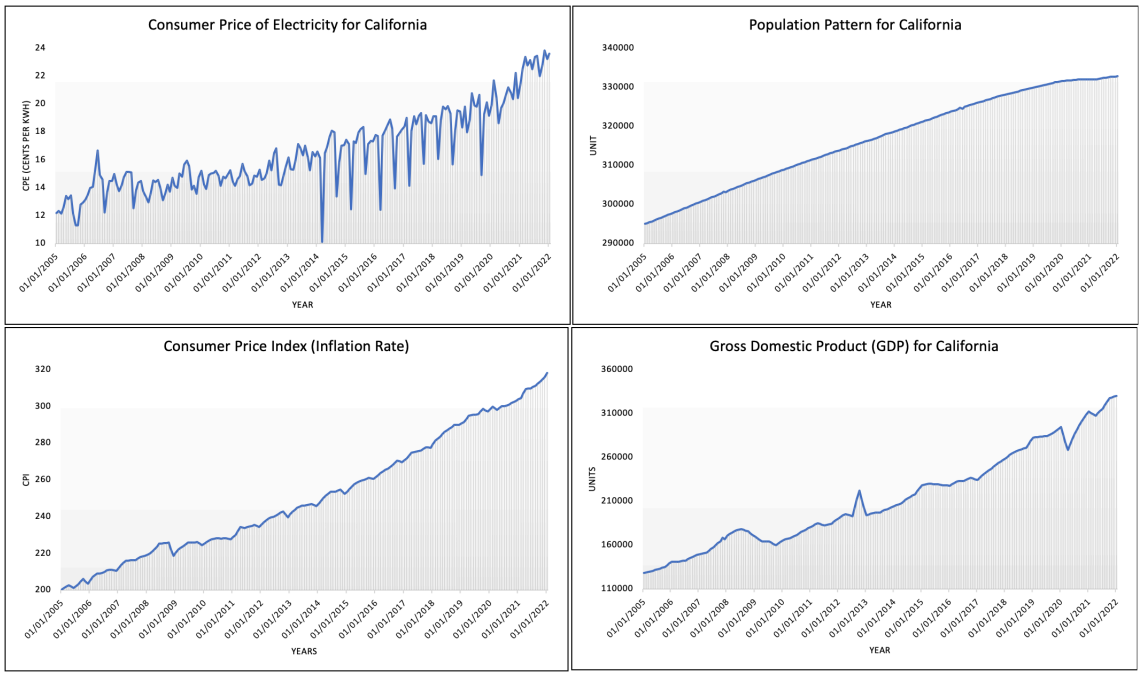

Fig. 4.2 The statistical representation of the dataset for independent variables.

4.3.3 Heat-map for the Correlation Matrix

In ML, feature selection is a method that considers only those independent features in the model that contribute significantly to estimating the dependent variable. Accordingly, the heat map of the correlation matrix is investigated to distinguish between the independent variables and the dependent variables. Figure 4.3 shows the heatmap for the parameters of the inflation rate, population, GDP, and CPE. The consumer price of electricity (CPE) and gross domestic product (GDP) has a correlation value of 0.91, indicating a strong relationship between these variables and their importance in estimating CPE. However, the GDP parameter exhibits

a correlation value lower than 0.85, suggesting a weaker relationship with CPE. In fact, including the population growth data led to no change in the final outcome of the results; however, it did lead to an increased computational time of the proposed model.

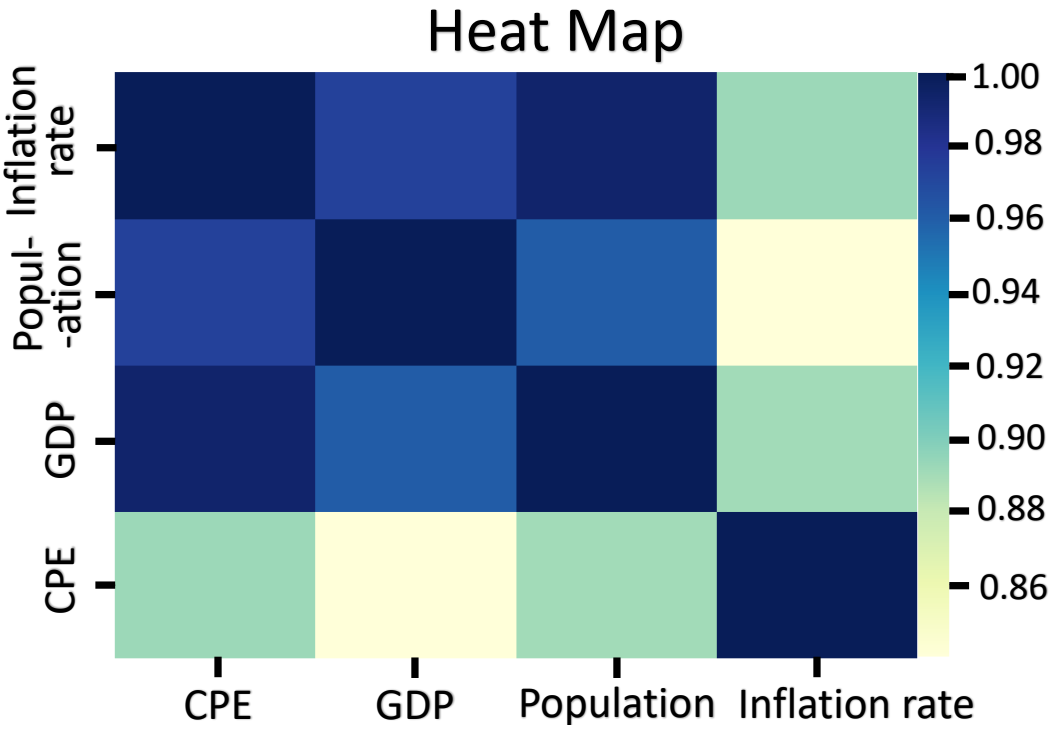

Fig. 4.3 The correlation matrix showcases the heatmap for evaluating the inter-dependency of variables concerning each other.

4.4 Results

This section discusses the implementation of various ML techniques using the aforementioned datasets and parameters to train and test the proposed model for accurately forecasting the LCOE and ROI of utility-based solar home systems. Additionally, the results are compared from the ML models with time series forecasting models such as ARIMA, LSTM, and Seasonal Autoregressive Integrated Moving Average (SARIMA). The results presented here focus on two locations: Sacramento, California, USA, and Butuan City, Philippines. These two locations are specifically chosen due to the availability of high-quality datasets for demographic variables. Furthermore, parameters such as CPE are consistent in these regions and do not vary based on rates determined by the government or industry. This consistency allows for a more reliable evaluation of the proposed model and its performance in predicting the LCOE and ROI for solar home systems.

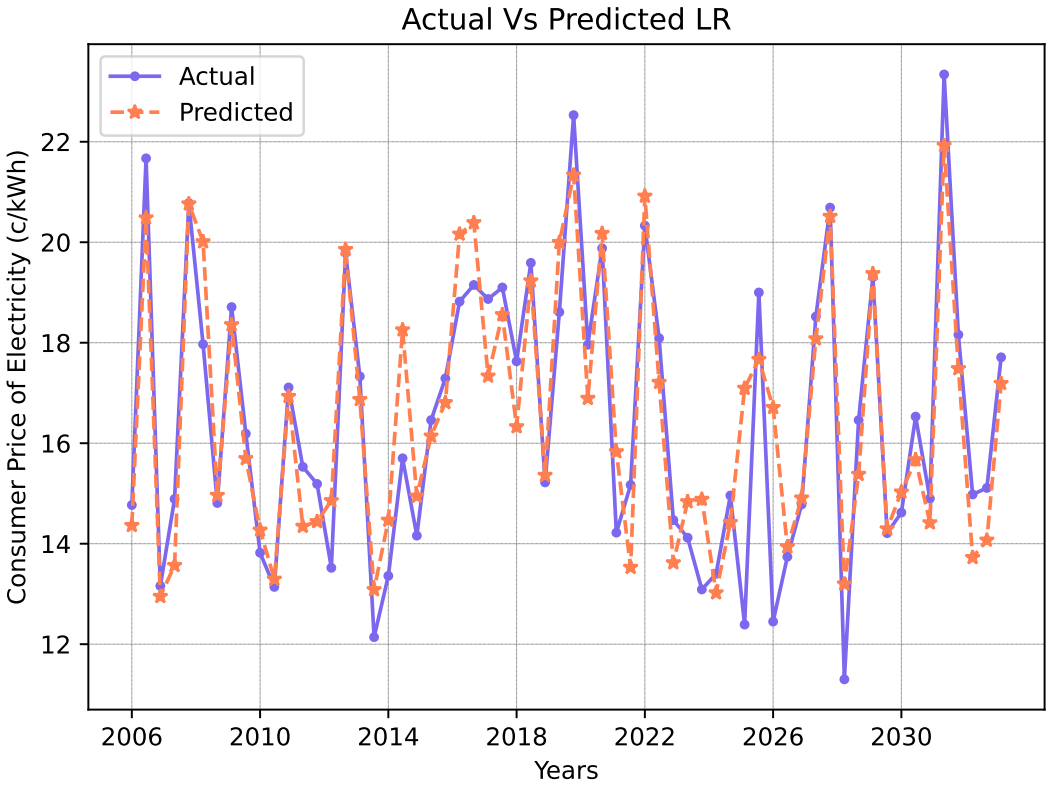

Fig. 4.4 The scattered plot of the predicted values to the actual values for the CPE (¢/kWh).

4.4.1 LR Model

For predicting the dependent variable CPE ($), the supervised learning ML model is applied initially and, precisely, the Linear Regression (LR) model. The LR model determines the best fit linear line between the independent and dependent variables. Moreover, the terms dependent and independent variables are already defined in the methodology section; subsequently, for implementing the LR model with input (dependent) variables such as population growth, the consumer price of index and GDP to calculate the dependent variable, i.e. CPE. Figure 4.4 shows the scattered plot of actual values concerning the predicted CPE in a linear relation. Accordingly, the LR model predicts the output values for the CPE over the next ten years.

The actual vs predicted plot for the CPE showed an accuracy of less than 85%, and the error loss function showed a root mean square error (RMSE) value of more than 10%. Therefore, according to the literature [165], the Accuracy should have a value of more than 90%, and RMSE should be less than 10% for the LR model to predict the values accurately. The limitation of such poor accuracy is that the input data for the independent variables was

limited and consisted of only three independent variables. Accordingly, good-quality data is used and added several independent variables.

4.4.2 LR Model with Multiple Variables

To improve the accuracy of the LR model, multiple independent variables are used and increased the duration of each variable, i.e. from January 2005 to December 2021. It is worth mentioning that there were some stances where the data was available in quarterly or annual resolutions. However, to enhance the accuracy of the LR model, the input data should be consistent and have the same time resolution. Accordingly, a tool called Generative adversarial networks (GAN) is used, a sub-class of ML in which two neural networks are considered.

Consequently, the GAN model results in the best ML model among these two neural networks. One of the advantages of the GAN model is to improve the quality of the model even with poor datasets. In addition, to predict the dependent variable CPE, the dataset is divided into 80% and 20% to train and test the LR model with multiple variables. Figure 4.5 showcases the results for the predicted values vs the actual values after executing the LR model with multiple variables. The overall accuracy for the LR model with multiple variables is 87%.

The accuracy achieved using the aforementioned model is within the limit of more than 85%. However, the model under consideration is not fruitful for accurately forecasting the LCOE and ROI parameters of the utility-based solar home system because it will still lead to ambiguity regarding the exact assumption of the ROI in terms of the year. Therefore, it is essential to appropriately determine an ML method with an accuracy of more than at least 90% [166]. In this regard, LSTM time series forecasting is incorporated to improve the accuracy of the ML model and reduce the loss error function.

4.4.3 LSTM Model

Another ML model, the Long Short-Term Memory (LSTM) model, was applied to enhance accuracy. The LSTM method belongs to a subset of ANNs within the domains of AI and deep neural networks. Additionally, the LSTM model is a Recurrent Neural Network (RNN) used for analyzing time series forecasting. Additionally, the model aims to predict the energy ROI, making time series forecasting crucial for accurately predicting the ROI of the installed system. Consequently, the results are extended by applying the LSTM Model with multiple independent variables.

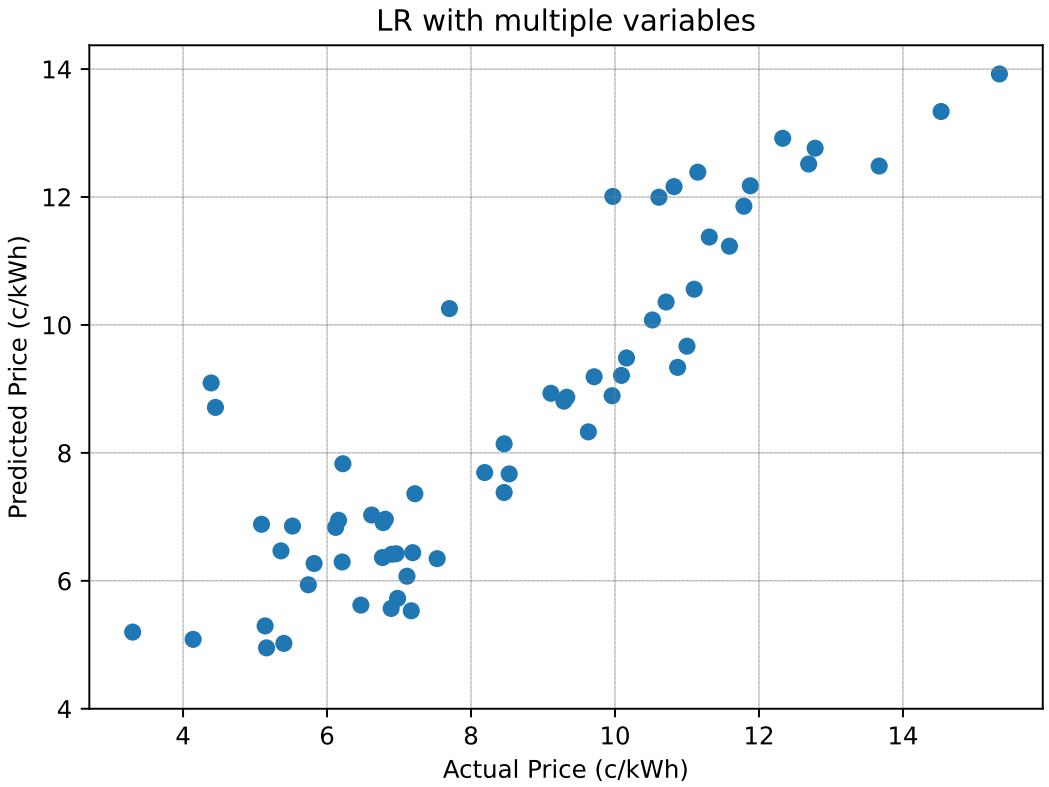

Fig. 4.5 The scattered plot shows the result of forecasting the dependent variable CPE (¢/kWh) using the LR model with Multiple input independent variables.

The results, demonstrated in Figure 4.6, show the dataset divided into 70% for training and 30% for testing. The LSTM model with multiple variables achieves an RMSE of 3.237% and an accuracy of 91%. The larger error in the early years of the LSTM model can be attributed to several factors. As mentioned earlier, the LSTM model is a type of RNN designed to capture long-term dependencies in sequential data. Initially, the model may struggle to capture these dependencies, leading to higher errors in the early stages of the time series. As the LSTM model progresses through the time series and continues to train, it gradually learns the underlying patterns and relationships in the data. This learning process enables the model to better capture long-term dependencies and adapt to the time series dynamics. Consequently, the model's predictions become more accurate over time, leading to a convergence of the error.

Figure 4.6, demonstrates the plot of the train and test of the LSTM model with multiple variables for predicting the dependent variable CPE (¢/kWh) for Sacramento, California, USA. Moreover, the predicted values of the CPE from the model are incorporated to calculate the LCOE and ROI of the utility-based solar home system. Though the LSTM model with

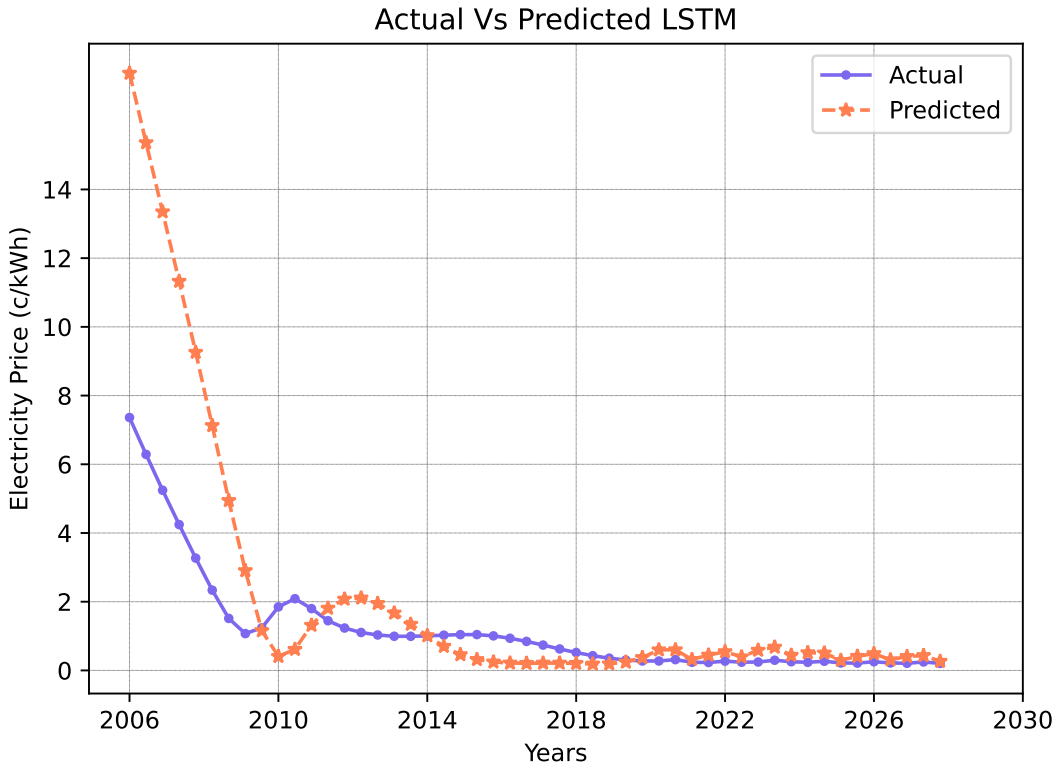

Fig. 4.6 The curve describes the plot of actual vs predicted values of the CPE (¢/kWh) using the LSTM model with multiple variables.

multiple variables achieved an accuracy of 91%, however, to obtain a more accurate model, the ARIMA model is further applied to test as discussed in the following subsection.

4.4.4 ARIMA Model

Next, the ARIMA model is applied to forecast the dependent variable, i.e. CPE. In general, an ARIMA model is a model that is fitted to the d^{th} order differenced time series that the resulting differenced time series needs to be stationary. Herein, the stationary time series is one in which the mean, variance, autocorrelation, and other statistical features remain constant across time. The ARIMA model is applied for time series forecasting for the study under consideration. In addition, it is worth mentioning that a similar dataset as an input is used to train and test the ARIMA model.

Apart from the high accuracy of the ARIMA model, there are several reasons for implementing it. First, ARIMA is a parametric model that offers interpretable coefficients that can be used to understand the underlying time series process. Second, the ARIMA model is highly flexible, as it can be applied to a wide range of time series data, including stationary, non-stationary, and seasonal data. Furthermore, ARIMA models can be extended to handle

exogenous variables, making them valuable in forecasting scenarios where other factors may impact the time series. Lastly, ARIMA models are robust to missing data and outliers, and numerous libraries and software packages provide built-in ARIMA functions.

Before applying the ARIMA model, the Dickey-Fuller algorithm is tested, a parameter to check the stationarity of the input dataset. The results of the Dickey-Fuller indicate if the dataset is stationary or not under the condition that the p-value (probability of the null hypothesis) should be very small. Accordingly, the results from the model gave a p-value of 0.23, which depicts that the dataset is stationary. In the ARIMA model, the AR part uses the previous values to make a future prediction, the MA uses the past errors for making future predictions, and the integrated here stands for the difference between the AR and MA.

Statistical tests were also incorporated to weigh each factor under consideration, such as the t-test and F-test, for assessing the significance of the individual coefficients of the AR, MA, and the constant term for each factor. The t-test determines the t-value for each coefficient, which expresses how far from zero the coefficient is in terms of standard errors. Indicating that the coefficient is statistically significant at the chosen level of significance (often 5% or 1%), a high t-value (generally larger than 2 or 2.5) is required. On the contrary, the combined significance of a set of model coefficients is evaluated using the F-test. The F-test is specifically used to test whether a subset of the coefficients—typically all the coefficients in a particular order—are equal to zero. When the p-value is low (often less than 0.05), the null hypothesis can be rejected and the subset of coefficients is jointly significant.

Furthermore, the model runs a set of interactions depending on the hit-and-trial method for calculating the most appropriate values for p (number of autoregressive terms), q (number of lagged forecast errors in the forecast equation) and d (number of nonseasonal differences required for stationarity). The results are analysed using Akaike's Information Criterion (AIC), which helps determine the predictors for the regression model. Subsequently, the model searches for the minimum AIC score and the (p,q, and d) values. Using these values as input data to the ARIMA model gave the minimum AIC score of 3214.29 and (p, q, d) values as (1, 0, 1), respectively.

In addition, the dataset was split into training (70%) and testing (30%) portions, along with the order (1, 0, 1) to apply the ARIMA model. Figure 4.7 demonstrates the actual (blue curve) vs predicted (orange curve) values for CPE, and the grey area highlights the confidence interval of 95% using these input values in the ARIMA model. It is worth mentioning that a confidence interval is a set of values surrounding a point estimate of a performance metric for a model (such as accuracy, precision, recall, etc.) that encapsulates the range of values in which the actual value of the performance metric is anticipated to reside with a given degree of confidence. The yellow-coloured dotted lines indicate the range of the predicted

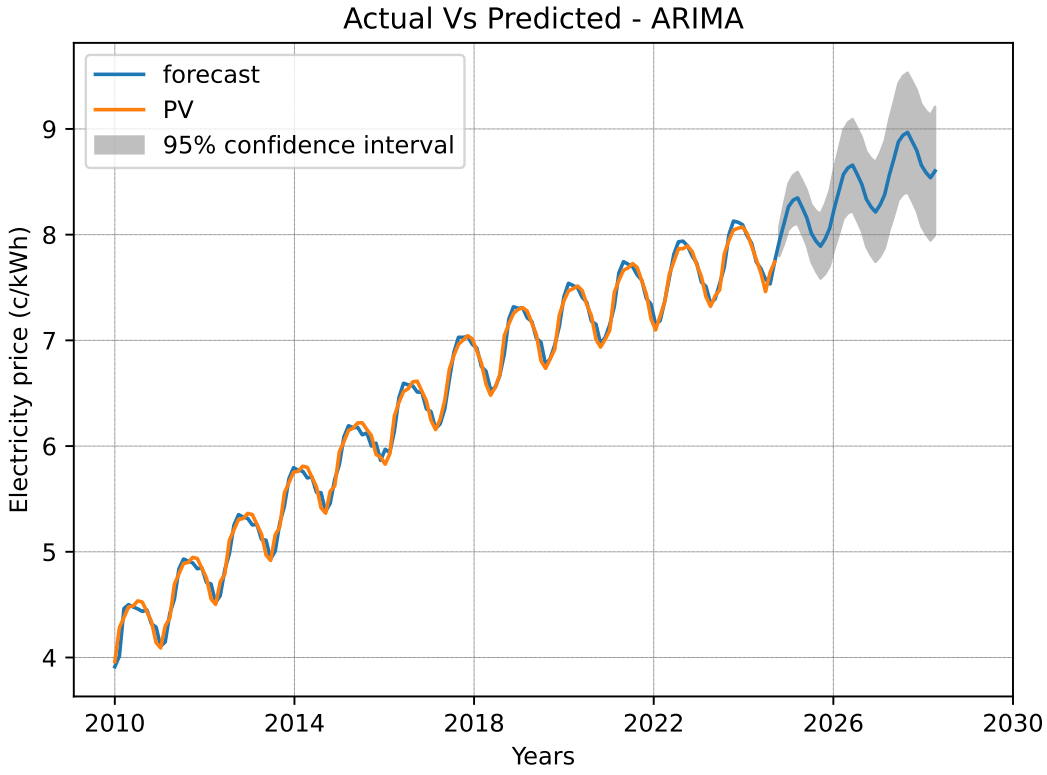

Fig. 4.7 The curve depicts the actual vs forecasted values for CPE (¢/kWh) for Sacramento, California, USA, using the ARIMA model.

values. Moreover, an accuracy of 93.8% is achieved and forecasted CPE values up to 2030. Therefore, the proposed model achieved a maximum of 93.8% accuracy and was the most appropriate model for predicting the LCOE parameter among other ML techniques.

Accordingly, after determining the most appropriate model for predicting one dependent variable, i.e. CPE. Consecutively, the same procedure is integrated for calculating the different other dependent variables (as mentioned in figure 4.1), such as solar panel cost ($), the balance of system cost ($), system lifetime (years), Operations and maintenance cost ($), Energy yield (($/kWh) and incentives ($). Consecutively, the dataset was collected for various independent variables, for example, type of solar panel, number of solar panels, the area required, the life span of devices, energy consumption, solar energy generated, associated breakdown costs, etc. From the literature, two case studies from Sacramento, California, USA and the Philippines were considered so as to make a detailed comparison of the proposed model. The dataset of the demographic variables ranges from a duration between 2005 to 2021.

4.5 Discussions

The results of applying ML for estimating the CPE show that the ARIMA model gave the highest accuracy. Accordingly, the independent variables for other parameters such as solar panel cost, the balance of system cost, system lifetime, operation and maintenance costs and the energy yield were forecasted accurately for evaluating the LCOE of a utility-based grid-connected solar home system. The dataset consisting of demographic variables is extracted for two countries, California, the USA and the Philippines and compared the results from the proposed model with the case studies. It is worth mentioning here that previously, the case studies calculating the LCOE used singular inputs rather than multiple variables, and an approach of ML is rarely shown in the literature. Therefore, the proposed model is the first approach to accurately forecasting the LCOE using the ML framework.

The first case study under consideration involves the overall project lifetime of 25 years for a grid-connected PV utility system of capacity 20MW installed in the city of Sacramento, California, USA. The overall performance of the system is 197 peak Watts per square meter. The estimated initial investment to install a PV system is $54 Million, and the contribution is $2.7/W having 65% modules and a 35% balance of systems. The direct purchase cost of the components involved in the PV utility system is $23.856 Million, and consecutively, the calculated values for the operation and maintenance cost of the PV system is shown in Table 4.2 with a total of $6.50 Million for 25 years.

These values are forecasted using the ARIMA model and ML techniques. The chosen PV module for the system is Monocrystalline-PERC (Passivated Emitter and Rear Cell), having an efficiency of 19.1%. Incorporating these input values into equations mentioned in the methodology section, the net energy production is calculated as 738.537GWh and the net present value of Electricity as 383.169GWh. The forecasted value of the Levelized total cost of electricity is $27.180 Million. Therefore, considering all the input values from the literature but integrating the values of CPE from the proposed model, the value of LCOE is obtained to be 7.09 ¢/kWh, whereas the LCOE using the singular inputs gives a value of 5.83 ¢/kWh. Similarly, using the equations mentioned in the methodology section, the forecasted ROI for the PV utility-based grid-connected system is 14 years.

In addition, to validate the proposed model, the dataset from another case study is used for a solar PV farm in a specific location in Butuan City, Philippines [167]. Similar to the previous, the initial dataset is extracted from the case study, such as the power capacity of the solar farm was 5 MW with an investment of 300 Million Pesos (to make the comparison with the first study, all the costs are converted to USD. Furthermore, the associated costs and the energy yield are included in calculating the solar farm's LCOE parameter and the ROI. According to the results of the case study for a duration of 20 years, the valuable energy

Table 4.2 The table showcases the calculation of the LCOE and the ROI of the PV system for the duration of 25 years.

Year	Production (GWh)	NPV of Electricity (GWh)	Direct Purchase Cost (M$)	Operation Maintenance (M$)	Levelized Total Cost ($)
0	–	–	23.856	–	23.856
1	31.343	29.569	–	0.26	0.245
2	31.187	27.757	–	0.26	0.231
3	31.032	26.055	–	0.26	0.218
4	30.878	24.458	–	0.26	0.206
5	30.724	22.959	–	0.26	0.194
6	30.571	21.552	–	0.26	0.183
7	30.419	20.231	–	0.26	0.173
8	30.268	18.990	–	0.26	0.163
9	30.117	17.826	–	0.26	0.154
10	29.967	16.734	–	0.26	0.145
11	29.818	15.708	–	0.26	0.137
12	29.670	14.745	–	0.26	0.129
13	29.522	13.841	–	0.26	0.122
14	29.376	12.993	–	0.26	0.115
15	29.229	12.196	–	0.26	0.108
16	29.084	11.449	–	0.26	0.102
17	28.939	10.747	–	0.26	0.097
18	28.795	10.088	–	0.26	0.091
19	28.652	9.470	–	0.26	0.086
20	28.509	8.889	–	0.26	0.081
21	28.368	8.344	–	0.26	0.076
22	28.226	7.833	–	0.26	0.072
23	28.086	7.353	–	0.26	0.068
24	27.946	6.902	–	0.26	0.064
25	27.807	6.479	–	0.26	0.061
Total	**738.537**	**383.169**	**23.856**	**6.50**	**27.180**

production is 4.18/kWh/day with an ROI of 4.23 years. However, their study used singular inputs to calculate these values. Consecutively, applying the ARIMA based-ML model in the proposed model, the resulting predicted value of the LCOE is 8.90¢/kWh while the ROI was calculated as 5.37 years.

Accordingly, analyzing the two case studies reveals that the demographic variables of any country will undoubtedly change over a period of time. Moreover, the discrepancy change in values using the singular inputs and the proposed model indicates that the LCOE and ROI calculated using the singular inputs results in errors and miscalculated estimation of the ROI of solar home systems.

4.6 Summary

In summary, the CPE represents the average cost of electricity per unit (in kWh) for consumers, which is influenced by various factors such as inflation rate, population growth, and gross domestic product. On the other hand, the LCOE is a metric that calculates the average cost of producing electricity per unit over the lifetime of a power generation system, such as a PV system. The relationship between CPE and LCOE in the proposed model is that the CPE serves as a reference point for estimating the LCOE of a PV system. In other words, CPE provides the context for comparing the cost-effectiveness of a PV system with the conventional prices of electricity. By considering the dynamic factors that influence CPE, the model aims to provide a more accurate estimation of LCOE, which ultimately helps in determining the ROI for installing a PV system. Incorporating the changing CPE in the proposed model allows for a more realistic and accurate assessment of the LCOE, which in turn contributes to a better understanding of the long-term financial viability of a PV system. Therefore, most of the existing studies rely on singular values.

However, the argument emphasized in this chapter investigates that many of these parameters are dynamic. Factors such as population growth, average past cost of electricity, inflation rate, gross domestic product, and other demographic variables significantly impact the cost of electricity. As a result, a model was developed that allows for the calculation of LCOE based on these dynamic input factors. Additionally, the results demonstrate a clear difference in estimating the LCOE of a PV system using singular inputs, yielding an LCOE of 5.83 ¢/kWh. In contrast, when applying the ML model, the LCOE increases to a value of 7.09 ¢/kWh. This comparison highlights the distinction between calculating the LCOE using singular inputs and employing ML and AI-based algorithms. Ultimately, the study in this chapter reveals a substantial difference in LCOE estimations, emphasizing the importance of considering dynamic factors.

Chapter 5

Conclusions and Future Work

5.1 Conclusion

In conclusion, the thesis presents a critical literature review, examining and implementing various ML techniques for predicting solar irradiance to enable WSNs. Lastly, the thesis also includes an ML-based framework for estimating the CPE of a PV system. Accordingly, a detailed conclusion of each of the chapters is presented below.

The literature review of recently published research articles was discussed that applied various ML techniques for predicting solar irradiance. Additionally, another section included the literature review for the ML algorithms that were previously used for the estimation of LCOE and ROI of PV systems. Moreover, an SR is also presented in the second chapter of the thesis, which covers a broad range of ML techniques for optimizing the performance of low-cost solar cells. The review indicates that a significant proportion of research focuses on data-driven approaches and ML techniques for discovering low-cost solar cells, with a third of publications targeting ML algorithms in the fabrication process. Moreover, the SR suggests that ML techniques can potentially accelerate the discovery of new solar materials and architectures. Future research can expand on these findings by exploring and developing new ML techniques for solar cell optimization. Additionally, addressing the scalability and sustainability of low-cost solar cell technologies to enable large-scale commercialization is essential. Ultimately, applying ML techniques in solar energy can revolutionize the industry and pave the way for a cleaner and more sustainable future.

Moreover, the WSN nodes rely on hazardous batteries that need constant replacement. Therefore, WSNs with solar energy harvesters that scavenge energy from the Sun are proposed. The critical issue with these harvesters is that solar energy is intermittent. Consequently, ML algorithms that enable WSN nodes to accurately predict the amount of solar irradiance are proposed so that the node can intelligently manage its energy. The ML models

were based on historical weather datasets from California (USA) and Delhi (India) from 2010 to 2020. In addition, data pre-processing, followed by feature engineering, identification of outliers and grid search to determine the most optimized ML model is performed. Compared with the linear regression model, the support vector regression (SVR) model showed accurate solar irradiance forecasting. Moreover, it was also found that the models with time duration of 1 year and 1 month has much better forecasting results than 10 years and 1 week, with both root square mean error (RMSE) and mean absolute error (MAE) less than 7% for Sacramento, California, USA.

In addition, most of the studies are based on singular values, and the argument here is that many of these parameters are dynamic. So, the parameters that determine the Cost of Electricity, like population growth, the average cost of electricity in the past, inflation rate, gross domestic product, and other demographic variables, have a large impact on the cost of electricity. Hence, a model that enables people to calculate the Levelized cost of electricity based on these dynamic input factors is developed. The compelling results show a clear difference in estimating the LCOE of a PV system using singular inputs; received the LCOE to be 5.83 ¢/kWh. However, applying the ML model, the LCOE is increased to 7.09 ¢/kWh. Thus, the study compares calculating the LCOE using the singular inputs and then the LCOE based on Machine learning and Artificial Intelligence based algorithms. Moreover, it signifies a big difference in estimating the LCOE values.

5.2 Open Questions

This section highlights some of the key insights and, consecutively, presents the future outlook of the potential research incorporating ML and the discovery of new materials to develop re-configurable solar cells. In addition, this section also includes the limitations and pitfalls of the ongoing research that needs to be addressed for developing efficient, robust, and stable solar cell architectures.

According to the review, few articles were published in the domain of using ML for fabricating solar cells. Furthermore, our study revealed that input data was clustered around PSCs, OSCs, and hybrid solar cells. Furthermore, most research used the ANN, GBRT, XGBoost, EXTR, LR, DTR, KNN, RF, SVM, SVR, GPR, and BO algorithms to determine output characteristics such as cost, PCE, the accuracy of the ML model, loss function and error. Lastly, ML was used for optimizing the following solar cell parameters: donor/acceptor ratio, conductivity, donor/acceptor materials, stability optimization, copper content optimization, and spray plasma processing.

Although there are numerous advantages of using ML for solar cell discovery, there are several open issues. From our systematic review, we came across multiple challenges that need to be addressed with regard to the discovery of new low-cost solar cells. Key among these challenges are:

- **Vulnerability of the input data**. As previously mentioned, most low-cost solar cells were fabricated by trial and error, which leads to high input data vulnerability [168]. Therefore, model validation should be a necessary step [169]. Moreover, data scarcity is a significant problem in the field of data-driven solar materials science [170]. Text mining and picture recognition are considered solutions to overcome these issues of small datasets [171].

- **Stability of thin-film solar cells**. One of the key concerns in designing low-cost solar cells in the real environment is the stability of organic, inorganic, and hybrid solar cells due to the different compositions of chemical components. These solar cells are very unstable and have a short life period [172]. Previously, studies have shown that solar cell efficiency and stability are inversely proportional. Also, the key stability components that need to be addressed are thermal, moisture, and chemical composition stability [173].

- **Inaccurate predictions**. Another key issue with using ML algorithms for discovering solar cells is the inaccurate predictions and outcomes from the ML models [174]. In most cases, ML algorithms give the confidence interval of the forecasted and predicted values of the solar cells. However, the predicted values for the discovery of solar cells seem to approach up to a maximum of 95% using the GPR and Bayesian optimization using the probability distribution, which sometimes proves to result in the poor fabrication of solar cells. Therefore, the ML models' prediction models need to be classified properly to avoid such discrepancies [175].

- **Rigorously fabricating solar cells in labs**. The researchers are rigorously fabricating solar cells depending upon the hit and trial methods, which wastes a lot of time, resources, and materials. In addition, if the researchers follow the same procedure in the upcoming years, it is noted that it will further delay the discovery of new materials used to fabricate solar cells [176]. Moreover, using the permutation and combinations of different layers, electrical characteristics, and other components required to design the solar cells and fabricate solar cells in the laboratory will lead to other consequences which can be avoided with the use of ML techniques and AI integration [177].

- **Lack of data availability and poor data analysis**. Firstly, it is noted from the study that there is a lack of data availability and, thus, poor data analysis. Second, it is advised to integrate feature engineering, modeling, and domain technical expertise to increase the effectiveness of the created ML model. In parallel, validation experiments should be run to verify the analytical outcomes of the ML model, such as the high-performing prediction candidate. Only a few research studies have used experiments to validate their forecasted materials [178].

5.3 Future Outlook

The future goals and prospective outlook for discovering new low-cost solar cells are mentioned below. Initially, there was a large room for data collection and monitoring to provide input to ML models. Moreover, the extracted data needs feature scaling and data-prepossessing to be used effectively in ML algorithms. Therefore, an appropriate data selection technique must be used to interpolate or extrapolate the data depending on various dependent and independent variables in feature selection. In addition, since ML and AI techniques have recently gained significant importance, adversarial robust ML techniques will play a vital role in forecasting and predicting the design of solar cell architectures.

Moreover, ML can aid in predicting the performance of solar cells, leading to the development of dependable and cost-effective solar cells. By predicting the performance of solar cells before production, manufacturers can save resources and avoid producing poorly performing cells. Additionally, ML is being utilized to create new materials for cost-effective solar cells. By analyzing large amounts of data from various sources, ML can identify materials with desired characteristics for solar cells, reducing the cost and time spent on experimentation and speeding up the process of developing new materials.

Since low-cost solar cell fabrication in a research laboratory is cheap, most researchers tend to retrospectively appreciate the performance of their design after first fabricating the solar cell by trial and error. Instead, we believe it is more beneficial to perform these predictions using robust ML algorithms, which will help design and fabricate more efficient solar cells. Adopting this approach will expedite the solar cell design process. There is also space for research related to the generalized explanations of data extraction and interpretation and to achieve more accurate ML models. In general, the accuracy of the ML model depends on the input data. Researchers across the globe should target to extract sufficient data and make it available online to help the scientific community discover low-cost, high-performance solar cells.

Bibliography

[1] Shouvik Mukherjee, Shariq Suleman, Roberto Pilloton, Jagriti Narang, and Kirti Rani. State of the art in smart portable, wearable, ingestible and implantable devices for health status monitoring and disease management. *Sensors*, 22(11):4228, 2022.

[2] Kui Zhao, Zhou Yang, and Shengzhong Liu. Emerging photovoltaic materials and devices, 2019.

[3] Yuchi Liu, Hamideh Khanbareh, Miah Abdul Halim, Andrew Feeney, Xiaosheng Zhang, Hadi Heidari, and Rami Ghannam. Piezoelectric energy harvesting for self-powered wearable upper limb applications. *Nano Select*, 2(8):1459–1479, 2021.

[4] Mahammad A Hannan, Saad Mutashar, Salina A Samad, and Aini Hussain. Energy harvesting for the implantable biomedical devices: issues and challenges. *Biomedical engineering online*, 13(1):1–23, 2014.

[5] Jinwei Zhao, Rami Ghannam, Kaung Oo Htet, Yuchi Liu, Man-kay Law, Vellaisamy AL Roy, Bruno Michel, Muhammad Ali Imran, and Hadi Heidari. Self-powered implantable medical devices: photovoltaic energy harvesting review. *Advanced healthcare materials*, 9(17):2000779, 2020.

[6] Rupam Das, Farshad Moradi, and Hadi Heidari. Biointegrated and wirelessly powered implantable brain devices: A review. *IEEE Transactions on Biomedical Circuits and Systems*, 14(2):343–358, 2020.

[7] Willem G De Voogt. Pacemaker leads: performance and progress. *The American journal of cardiology*, 83(5):187–191, 1999.

[8] Eric Stach, Brian DeCost, A Gilad Kusne, Jason Hattrick-Simpers, Keith A Brown, Kristofer G Reyes, Joshua Schrier, Simon Billinge, Tonio Buonassisi, Ian Foster, et al. Autonomous experimentation systems for materials development: A community perspective. *Matter*, 4(9):2702–2726, 2021.

[9] Maram A Wahba, Amira S Ashour, and Rami Ghannam. Prediction of harvestable energy for self-powered wearable healthcare devices: Filling a gap. *IEEE Access*, 8:170336–170354, 2020.

[10] Rishi E Kumar, Armi Tiihonen, Shijing Sun, David P Fenning, Zhe Liu, and Tonio Buonassisi. Opportunities for machine learning to accelerate halide-perovskite commercialization and scale-up. *Matter*, 5(5):1353–1366, 2022.

[11] PTV Bhuvaneswari, R Balakumar, Vijay Vaidehi, and P Balamuralidhar. Solar energy harvesting for wireless sensor networks. In *2009 First International Conference on Computational Intelligence, Communication Systems and Networks*, pages 57–61. IEEE, 2009.

[12] Mitch Jacoby. The future of low-cost solar cells. *Chem. Eng. News*, 94(18):30–35, 2016.

[13] Matthew C Beard, Joseph M Luther, and Arthur J Nozik. The promise and challenge of nanostructured solar cells. *Nature nanotechnology*, 9(12):951–954, 2014.

[14] Rami Ghannam, Paulo Valente Klaine, and Muhammad Imran. Artificial intelligence for photovoltaic systems. In *Power Systems*, pages 121–142. Springer Singapore, 2019.

[15] Hong Duc Pham, Zhifang Wu, Luis K Ono, Sergei Manzhos, Krishna Feron, Nunzio Motta, Yabing Qi, and Prashant Sonar. Low-cost alternative high-performance hole-transport material for perovskite solar cells and its comparative study with conventional spiro-ometad. *Advanced Electronic Materials*, 3(8):1700139, 2017.

[16] Çağla Odabaşı and Ramazan Yıldırım. Performance analysis of perovskite solar cells in 2013–2018 using machine-learning tools. *Nano Energy*, 56:770–791, 2019.

[17] Asif Mahmood, Jing Yang, Junyi Hu, Xiaochen Wang, Ailing Tang, Yanfang Geng, Qingdao Zeng, and Erjun Zhou. Introducing four 1, 1-dicyanomethylene-3-indanone end-capped groups as an alternative strategy for the design of small-molecular non-fullerene acceptors. *The Journal of Physical Chemistry C*, 122(51):29122–29128, 2018.

[18] Huaxing Zhou, Liqiang Yang, Samuel C Price, Kelly Jane Knight, and Wei You. Enhanced photovoltaic performance of low-bandgap polymers with deep lumo levels. *Angewandte chemie*, 122(43):8164–8167, 2010.

[19] Feng Chen, Mei-Hong Liu, Rui-Qi Piao, De-Long Zhang, and Yan Wang. Cross-section spectra and transient characteristics of er3+ emissions in gd3 (al, ga) 5o12 garnet single crystal. *Optical Materials*, 136:113439, 2023.

[20] Issam El Naqa and Martin J Murphy. What is machine learning? In *machine learning in radiation oncology*, pages 3–11. Springer, 2015.

[21] H Wang, ZeZXeZBePJ Lei, X Zhang, B Zhou, and J Peng. Machine learning basics. *Deep Learn*, pages 98–164, 2016.

[22] Jude W Shavlik, Thomas Dietterich, and Thomas Glen Dietterich. *Readings in machine learning*. Morgan Kaufmann, 1990.

[23] Foster Provost and Ron Kohavi. On applied research in machine learning. *MACHINE LEARNING-BOSTON-*, 30:127–132, 1998.

[24] Asif Mahmood, Junyi Hu, Ailing Tang, Fan Chen, Xiaochen Wang, and Erjun Zhou. A novel thiazole based acceptor for fullerene-free organic solar cells. *Dyes and Pigments*, 149:470–474, 2018.

[25] Michael I Jordan and Tom M Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.

[26] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[27] David H Wolpert and William G Macready. No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, 1(1):67–82, 1997.

[28] Felipe Oviedo, Zekun Ren, Xue Hansong, Siyu Isaac Parker Tian, Kaicheng Zhang, Mariya Layurova, Thomas Heumueller, Ning Li, Erik Birgersson, Shijing Sun, et al. Bridging the gap between photovoltaics r&d and manufacturing with data-driven optimization. *arXiv preprint arXiv:2004.13599*, 2020.

[29] Asif Mahmood, Ahmad Irfan, and Jin-Liang Wang. Machine learning for organic photovoltaic polymers: A minireview. *Chinese Journal of Polymer Science*, pages 1–7, 2022.

[30] Lei Zhang, Mu He, and Shaofeng Shao. Machine learning for halide perovskite materials. *Nano Energy*, 78:105380, 2020.

[31] Keith T Butler, Daniel W Davies, Hugh Cartwright, Olexandr Isayev, and Aron Walsh. Machine learning for molecular and materials science. *Nature*, 559(7715):547–555, 2018.

[32] Nishi Parikh, Meera Karamta, Neha Yadav, Mohammad Mahdi Tavakoli, Daniel Prochowicz, Seckin Akin, Abul Kalam, Soumitra Satapathi, and Pankaj Yadav. Is machine learning redefining the perovskite solar cells? *Journal of Energy Chemistry*, 66:74–90, 2022.

[33] Maniell Workman, David Zhi Chen, and Sarhan M Musa. Machine learning for predicting perovskite solar cell opto-electronic properties. In *2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI)*, pages 1–5. IEEE, 2020.

[34] Harsh Rajesh Parikh, Yoann Buratti, Sergiu Spataru, Frederik Villebro, Gisele Alves Dos Reis Benatto, Peter B Poulsen, Stefan Wendlandt, Tamas Kerekes, Dezso Sera, and Ziv Hameiri. Solar cell cracks and finger failure detection using statistical parameters of electroluminescence images and machine learning. *Applied Sciences*, 10(24):8834, 2020.

[35] Daniil Bash, Frederick Hubert Chenardy, Zekun Ren, Jayce J Cheng, Tonio Buonassisi, Ricardo Oliveira, Jatin N Kumar, and Kedar Hippalgaonkar. Accelerated automated screening of viscous graphene suspensions with various surfactants for optimal electrical conductivity. *Digital Discovery*, 1(2):139–146, 2022.

[36] Qiuling Tao, Pengcheng Xu, Minjie Li, and Wencong Lu. Machine learning for perovskite materials design and discovery. *npj Computational Materials*, 7(1):1–18, 2021.

[37] Hannes Michaels, Iacopo Benesperi, and Marina Freitag. Challenges and prospects of ambient hybrid solar cell applications. *Chemical Science*, 12(14):5002–5015, 2021.

[38] Hannes Wagner-Mohnsen and Pietro P Altermatt. Machine learning for optimization of mass-produced industrial silicon solar cells. In *2021 International Conference on Numerical Simulation of Optoelectronic Devices (NUSOD)*, pages 51–52. IEEE, 2021.

[39] Yongjie Cui, Peipei Zhu, Xunfan Liao, and Yiwang Chen. Recent advances of computational chemistry in organic solar cell research. *Journal of Materials Chemistry C*, 8(45):15920–15939, 2020.

[40] Florian Häse, Loïc M Roch, Pascal Friederich, and Alán Aspuru-Guzik. Designing and understanding light-harvesting devices with machine learning. *Nature Communications*, 11(1):1–11, 2020.

[41] Sheng Jiang, Cun-Cun Wu, Fan Li, Yu-Qing Zhang, Ze-Hao Zhang, Qiao-Hui Zhang, Zhi-Jian Chen, Bo Qu, Li-Xin Xiao, and Min-Lin Jiang. Machine learning (ml)-assisted optimization doping of ki in mapbi3 solar cells. *Rare Metals*, 40(7):1698–1707, 2021.

[42] Anton O Oliynyk and Jillian M Buriak. Virtual issue on machine-learning discoveries in materials science, 2019.

[43] Min-Hsuan Lee. Performance and matching band structure analysis of tandem organic solar cells using machine learning approaches. *Energy Technology*, 8(3):1900974, 2020.

[44] Amandeep Sharma and Ajay Kakkar. Machine learning based optimal renewable energy allocation in sustained wireless sensor networks. *Wireless Networks*, 25(7):3953–3981, 2019.

[45] Amandeep Sharma and Ajay Kakkar. Forecasting daily global solar irradiance generation using machine learning. *Renewable and Sustainable Energy Reviews*, 82:2254–2269, 2018.

[46] Sukham Dhillon, Charu Madhu, Daljeet Kaur, and Sarvjit Singh. A solar energy forecast model using neural networks: Application for prediction of power for wireless sensor networks in precision agriculture. *Wireless Personal Communications*, 112:2741–2760, 2020.

[47] Muhammad Faizan Ghuman, Adnan Iqbal, Hassaan Khaliq Qureshi, and Marios Lestas. Asim: Solar energy availability model for wireless sensor networks. In *Proceedings of the 3rd International Workshop on Energy Harvesting & Energy Neutral Sensing Systems*, pages 21–26, 2015.

[48] Kadra Branker, MJM Pathak, and Joshua M Pearce. A review of solar photovoltaic levelized cost of electricity. *Renewable and sustainable energy reviews*, 15(9):4470–4482, 2011.

[49] Chul-Yong Lee and Jaekyun Ahn. Stochastic modeling of the levelized cost of electricity for solar pv. *Energies*, 13(11):3017, 2020.

[50] Ahsan Raza Khan, Sohail Razzaq, Thamer Alquthami, Muhammad Riaz Moghal, Adil Amin, and Anzar Mahmood. Day ahead load forecasting for iesco using artificial neural network and bagged regression tree. In *2018 1st International Conference on Power, Energy and Smart Grid (ICPESG)*, pages 1–6. IEEE, 2018.

[51] Thomas Geissmann and Oriana Ponta. A probabilistic approach to the computation of the levelized cost of electricity. *Energy*, 124:372–381, 2017.

[52] T Georgitsioti, N Pearsall, and I Forbes. The simplified levelized cost of the domestic pv energy in the uk: The importance of the feed-in tariff scheme. *Proc. Photovoltaic Science, Applications and Technology (PVSAT-9)*, pages 9–12, 2013.

[53] Jino Im, Seongwon Lee, Tae-Wook Ko, Hyun Woo Kim, YunKyong Hyon, and Hyunju Chang. Identifying pb-free perovskites for solar cells by machine learning. *npj Computational Materials*, 5(1):1–8, 2019.

[54] Caroline Persson, Peder Bacher, Takahiro Shiga, and Henrik Madsen. Multi-site solar power forecasting using gradient boosted regression trees. *Solar Energy*, 150:423–436, 2017.

[55] Jinxin Li, Basudev Pradhan, Surya Gaur, and Jayan Thomas. Predictions and strategies learned from machine learning to develop high-performing perovskite solar cells. *Advanced Energy Materials*, 9(46):1901891, 2019.

[56] Pavlos Nikolaidis and Sotirios Chatzis. Gaussian process-based bayesian optimization for data-driven unit commitment. *International Journal of Electrical Power & Energy Systems*, 130:106930, 2021.

[57] Maniell Workman, David Zhi Chen, and Sarhan M. Musa. Machine learning for predicting perovskite solar cell opto-electronic properties. In *2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI)*, pages 1–5, 2020.

[58] Zhi Li, Mansoor Ani Najeeb, Liana Alves, Alyssa Z Sherman, Venkateswaran Shekar, Peter Cruz Parrilla, Ian M Pendleton, Wesley Wang, Philip W Nega, Matthias Zeller, et al. Robot-accelerated perovskite investigation and discovery. *Chemistry of Materials*, 32(13):5650–5663, 2020.

[59] Chenglong She, Qicheng Huang, Cong Chen, Yue Jiang, Zhen Fan, and Jinwei Gao. Machine learning-guided search for high-efficiency perovskite solar cells with doped electron transport layers. *Journal of Materials Chemistry A*, 9(44):25168–25177, 2021.

[60] Hong-Jian Feng and Ping Ma. Machine learning prediction of 2d perovskite photovoltaics and interaction with energetic ion implantation. *Applied Physics Letters*, 119(23):231902, 2021.

[61] Elif Ceren Gok, Murat Onur Yildirim, Muhammed PU Haris, Esin Eren, Meenakshi Pegu, Naveen Harindu Hemasiri, Peng Huang, Samrana Kazim, Aysegul Uygun Oksuz, and Shahzada Ahmad. Predicting perovskite bandgap and solar cell performance with machine learning. *Solar RRL*, 6(2):2100927, 2022.

[62] Xia Cai, Fengcai Liu, Anran Yu, Jiajun Qin, Mohammad Hatamvand, Irfan Ahmed, Jiayan Luo, Yiming Zhang, Hao Zhang, and Yiqiang Zhan. Data-driven design of high-performance masnxpb1-xi3 perovskite materials by machine learning and experimental realization. *Light: Science & Applications*, 11(1):1–12, 2022.

[63] Shinji Nagasawa, Eman Al-Naamani, and Akinori Saeki. Computer-aided screening of conjugated polymers for organic solar cell: classification by random forest. *The Journal of Physical Chemistry Letters*, 9(10):2639–2646, 2018.

[64] Harikrishna Sahu, Weining Rao, Alessandro Troisi, and Haibo Ma. Toward predicting efficiency of organic solar cells via machine learning and improved descriptors. *Advanced Energy Materials*, 8(24):1801032, 2018.

[65] Harikrishna Sahu and Haibo Ma. Unraveling correlations between molecular properties and device parameters of organic solar cells using machine learning. *The journal of physical chemistry letters*, 10(22):7277–7284, 2019.

[66] Daniele Padula, Jack D Simpson, and Alessandro Troisi. Combining electronic and structural features in machine learning models to predict organic solar cells properties. *Materials Horizons*, 6(2):343–349, 2019.

[67] Min-Hsuan Lee. Robust random forest based non-fullerene organic solar cells efficiency prediction. *Organic Electronics*, 76:105465, 2020.

[68] Xiaoyan Du, Larry Lüer, Thomas Heumueller, Jerrit Wagner, Christian Berger, Tobias Osterrieder, Jonas Wortmann, Stefan Langner, Uyxing Vongsaysy, Melanie Bertrand, et al. Elucidating the full potential of opv materials utilizing a high-throughput robot-based platform and machine learning. *Joule*, 5(2):495–506, 2021.

[69] Ahmad Irfan, Mohamed Hussien, Muhammad Yasir Mehboob, Aziz Ahmad, and Muhammad Ramzan Saeed Ashraf Janjua. Learning from fullerenes and predicting for y6: Machine learning and high-throughput screening of small molecule donors for organic solar cells. *Energy Technology*, page 2101096, 2022.

[70] Wei Chen, Yongzhen Wu, Youfeng Yue, Jian Liu, Wenjun Zhang, Xudong Yang, Han Chen, Enbing Bi, Islam Ashraful, Michael Grätzel, et al. Efficient and stable large-area perovskite solar cells with inorganic charge extraction layers. *Science*, 350(6263):944–948, 2015.

[71] Asif Mahmood and Jin-Liang Wang. Machine learning for high performance organic solar cells: current scenario and future prospects. *Energy & environmental science*, 14(1):90–105, 2021.

[72] Lingfei Wei, Xiaojie Xu, Gurudayal, James Bullock, and Joel W Ager. Machine learning optimization of p-type transparent conducting films. *Chemistry of materials*, 31(18):7340–7350, 2019.

[73] Junjie Peng, Elizabeth C Jury, Pierre Dönnes, and Coziana Ciurtin. Machine learning techniques for personalised medicine approaches in immune-mediated chronic inflammatory diseases: applications and challenges. *Frontiers in Pharmacology*, page 2667, 2021.

[74] Atieh Hashemi, Majid Basafa, and Aidin Behravan. Machine learning modeling for solubility prediction of recombinant antibody fragment in four different e. coli strains. *Scientific reports*, 12(1):1–11, 2022.

[75] Yiming Liu, Wensheng Yan, Heng Zhu, Yiteng Tu, Li Guan, and Xinyu Tan. Study on bandgap predications of abx3-type perovskites by machine learning. *Organic Electronics*, 101:106426, 2022.

[76] Lei Zhang and Mu He. Prediction of solar cell materials via unsupervised literature learning. *Journal of Physics: Condensed Matter*, 34(9):095902, 2021.

[77] Ian Mathews, Sai Nithin Reddy Kantareddy, Shijing Sun, Mariya Layurova, Janak Thapa, Juan-Pablo Correa-Baena, Rahul Bhattacharyya, Tonio Buonassisi, Sanjay Sarma, and Ian Marius Peters. Self-powered sensors enabled by wide-bandgap perovskite indoor photovoltaic cells. *Advanced Functional Materials*, 29(42):1904072, 2019.

[78] Jason J Yoo, Sarah Wieghold, Melany C Sponseller, Matthew R Chua, Sophie N Bertram, Noor Titan Putri Hartono, Jason S Tresback, Eric C Hansen, Juan-Pablo Correa-Baena, Vladimir Bulović, et al. An interface stabilized perovskite solar cell with high stabilized efficiency and low voltage loss. *Energy & Environmental Science*, 12(7):2192–2199, 2019.

[79] Ian Mathews, Sai Nithin Kantareddy, Tonio Buonassisi, and Ian Marius Peters. Technology and market perspective for indoor photovoltaic cells. *Joule*, 3(6):1415–1426, 2019.

[80] Tianmin Wu and Jian Wang. Deep mining stable and nontoxic hybrid organic–inorganic perovskites for photovoltaics via progressive machine learning. *ACS Applied Materials & Interfaces*, 12(52):57821–57831, 2020.

[81] Zhi-Wen Zhao, Marcos del Cueto, Yun Geng, and Alessandro Troisi. Effect of increasing the descriptor set on machine learning prediction of small molecule-based organic solar cells. *Chemistry of Materials*, 32(18):7777–7787, 2020.

[82] Yao Wu, Jie Guo, Rui Sun, and Jie Min. Machine learning for accelerating the discovery of high-performance donor/acceptor pairs in non-fullerene organic solar cells. *npj Computational Materials*, 6(1):1–8, 2020.

[83] Kakaraparthi Kranthiraja and Akinori Saeki. Experiment-oriented machine learning of polymer: non-fullerene organic solar cells. *Advanced Functional Materials*, 31(23):2011168, 2021.

[84] Shijing Sun, Armi Tiihonen, Felipe Oviedo, Zhe Liu, Janak Thapa, Yicheng Zhao, Noor Titan P Hartono, Anuj Goyal, Thomas Heumueller, Clio Batali, et al. A data fusion approach to optimize compositional stability of halide perovskites. *Matter*, 4(4):1305–1322, 2021.

[85] Tianmin Wu and Jian Wang. Global discovery of stable and non-toxic hybrid organic-inorganic perovskites for photovoltaic systems by combining machine learning method with first principle calculations. *Nano Energy*, 66:104070, 2019.

[86] Oleksandr Voznyy, Larissa Levina, James Z Fan, Mikhail Askerka, Ankit Jain, Min-Jae Choi, Olivier Ouellette, Petar Todorovic, Laxmi K Sagar, and Edward H Sargent. Machine learning accelerates discovery of optimal colloidal quantum dot synthesis. *ACS nano*, 13(10):11122–11128, 2019.

[87] Michael L. Agiorgousis, Yi-Yang Sun, Duk-Hyun Choe, Damien West, and Shengbai Zhang. Machine learning augmented discovery of chalcogenide double perovskites for photovoltaics. *Advanced Theory and Simulations*, 2(5):1800173, 2019.

[88] Nastaran Meftahi, Mykhailo Klymenko, Andrew J Christofferson, Udo Bach, David A Winkler, and Salvy P Russo. Machine learning property prediction for organic photovoltaic devices. *npj Computational Materials*, 6(1):1–8, 2020.

[89] Noor Titan Putri Hartono, Janak Thapa, Armi Tiihonen, Felipe Oviedo, Clio Batali, Jason J Yoo, Zhe Liu, Ruipeng Li, David Fuertes Marrón, Moungi G Bawendi, et al. How machine learning can help select capping layers to suppress perovskite degradation. *Nature communications*, 11(1):1–9, 2020.

[90] Zhilong Wang, Haikuo Zhang, and Jinjin Li. Accelerated discovery of stable spinels in energy systems via machine learning. *Nano Energy*, 81:105665, 2021.

[91] Muhammad Ramzan Saeed Ashraf Janjua, Ahmad Irfan, Mohamed Hussien, Muhammad Ali, Muhammad Saqib, and Muhammad Sulaman. Machine-learning analysis of small-molecule donors for fullerene based organic solar cells. *Energy Technology*, 10(5):2200019, 2022.

[92] Harikrishna Sahu, Feng Yang, Xiaobo Ye, Jing Ma, Weihai Fang, and Haibo Ma. Designing promising molecules for organic solar cells via machine learning assisted virtual screening. *Journal of Materials Chemistry A*, 7(29):17480–17488, 2019.

[93] Shohei Kanno, Yutaka Imamura, and Masahiko Hada. Alternative materials for perovskite solar cells from materials informatics. *Physical Review Materials*, 3(7):075403, 2019.

[94] Lifei Ju, Minjie Li, Lumin Tian, Pengcheng Xu, and Wencong Lu. Accelerated discovery of high-efficient n-annulated perylene organic sensitizers for solar cells via machine learning and quantum chemistry. *Materials Today Communications*, 25:101604, 2020.

[95] Wissam A Saidi, Waseem Shadid, and Ivano E Castelli. Machine-learning structural and electronic properties of metal halide perovskites using a hierarchical convolutional neural network. *npj Computational Materials*, 6(1):1–7, 2020.

[96] Yaping Wen, Lulu Fu, Gongqiang Li, Jing Ma, and Haibo Ma. Accelerated discovery of potential organic dyes for dye-sensitized solar cells by interpretable machine learning models and virtual screening. *Solar RRL*, 4(6):2000110, 2020.

[97] Zongmei Guo and Bin Lin. Machine learning stability and band gap of lead-free halide double perovskite materials for perovskite solar cells. *Solar Energy*, 228:689–699, 2021.

[98] Jialu Chen, Wenjun Xu, and Ruiqin Zhang. δ-machine learning-driven discovery of double hybrid organic–inorganic perovskites. *Journal of Materials Chemistry A*, 10(3):1402–1413, 2022.

[99] John M Howard, Elizabeth M Tennyson, Bernardo RA Neves, and Marina S Leite. Machine learning for perovskites' reap-rest-recovery cycle. *Joule*, 3(2):325–337, 2019.

[100] Jeffrey Kirman, Andrew Johnston, Douglas A Kuntz, Mikhail Askerka, Yuan Gao, Petar Todorović, Dongxin Ma, Gilbert G Privé, and Edward H Sargent. Machine-learning-accelerated perovskite crystallization. *Matter*, 2(4):938–947, 2020.

[101] Lei Zhang and Mu He. Unsupervised machine learning for solar cell materials from the literature. *Journal of Applied Physics*, 131(6):064902, 2022.

[102] Asif Mahmood, Ahmad Irfan, and Jin-Liang Wang. Machine learning and molecular dynamics simulation-assisted evolutionary design and discovery pipeline to screen efficient small molecule acceptors for ptb7-th-based organic solar cells with over 15% efficiency. *Journal of Materials Chemistry A*, 10(8):4170–4180, 2022.

[103] Bing Cao, Lawrence A Adutwum, Anton O Oliynyk, Erik J Luber, Brian C Olsen, Arthur Mar, and Jillian M Buriak. How to optimize materials and devices via design of experiments and machine learning: Demonstration using organic photovoltaics. *ACS nano*, 12(8):7434–7444, 2018.

[104] Asif Mahmood, Ahmad Irfan, and Jin-Liang Wang. Machine learning for organic photovoltaic polymers: A minireview. *Chinese Journal of Polymer Science*, 40(8):870–876, 2022.

[105] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[106] Asif Mahmood, Ahmad Irfan, and Jin-Liang Wang. Developing efficient small molecule acceptors with sp2-hybridized nitrogen at different positions by density functional theory calculations, molecular dynamics simulations and machine learning. *Chemistry–A European Journal*, 28(2):e202103712, 2022.

[107] Daniele Padula and Alessandro Troisi. Concurrent optimization of organic donor–acceptor pairs through machine learning. *Advanced Energy Materials*, 9(40):1902463, 2019.

[108] Wenbo Sun, Yujie Zheng, Ke Yang, Qi Zhang, Akeel A Shah, Zhou Wu, Yuyang Sun, Liang Feng, Dongyang Chen, Zeyun Xiao, et al. Machine learning–assisted molecular design and efficiency prediction for high-performance organic photovoltaic materials. *Science advances*, 5(11):eaay4275, 2019.

[109] Tudur Wyn David, Helder Anizelli, T Jesper Jacobsson, Cameron Gray, William Teahan, and Jeff Kettle. Enhancing the stability of organic photovoltaics through machine learning. *Nano Energy*, 78:105342, 2020.

[110] Tudur Wyn David, Helder Anizelli, Priyanka Tyagi, Cameron Gray, William Teahan, and Jeff Kettle. Using large datasets of organic photovoltaic performance data to elucidate trends in reliability between 2009 and 2019. *IEEE Journal of Photovoltaics*, 9(6):1768–1773, 2019.

[111] Asif Mahmood and Jin-Liang Wang. A time and resource efficient machine learning assisted design of non-fullerene small molecule acceptors for p3ht-based organic solar cells and green solvent selection. *Journal of Materials Chemistry A*, 9(28):15684–15695, 2021.

[112] Ghaith Salman, Stephen M Goodnick, Abdul R Shaik, and Dragica Vasileska. Machine learning for optimal copper doping profile design in cdte solar cells. In *2021 IEEE 48th Photovoltaic Specialists Conference (PVSC)*, pages 0540–0543. IEEE, 2021.

[113] Mona Gafar, Ragab A El-Sehiemy, Hany M Hasanien, and Amlak Abaza. Optimal parameter estimation of three solar cell models using modified spotted hyena optimization. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–12, 2022.

[114] Julia WP Hsu and Weijie Xu. Accelerate process optimization in perovskite solar cell manufacturing with machine learning. *Matter*, 5(5):1334–1336, 2022.

[115] Michael A Gelbart, Jasper Snoek, and Ryan P Adams. Bayesian optimization with unknown constraints. *arXiv preprint arXiv:1403.5607*, 2014.

[116] Zhe Liu, Nicholas Rolston, Austin C Flick, Thomas W Colburn, Zekun Ren, Reinhold H Dauskardt, and Tonio Buonassisi. Machine learning with knowledge constraints for process optimization of open-air perovskite solar cell manufacturing. *Joule*, 6(4):834–849, 2022.

[117] Vincent M Le Corre, Tejas S Sherkar, Marten Koopmans, and L Jan Anton Koster. Identification of the dominant recombination process for perovskite solar cells based on machine learning. *Cell Reports Physical Science*, 2(2):100346, 2021.

[118] Xabier Rodríguez-Martínez, Enrique Pascual-San-José, and Mariano Campoy-Quiles. Accelerating organic solar cell material's discovery: high-throughput screening and big data. *Energy & Environmental Science*, 14(6):3301–3322, 2021.

[119] Aaron Kirkey, Erik J Luber, Bing Cao, Brian C Olsen, and Jillian M Buriak. Optimization of the bulk heterojunction of all-small-molecule organic photovoltaics using design of experiment and machine learning approaches. *ACS Applied Materials & Interfaces*, 12(49):54596–54607, 2020.

[120] Çağla Odabaşı and Ramazan Yıldırım. Machine learning analysis on stability of perovskite solar cells. *Solar Energy Materials and Solar Cells*, 205:110284, 2020.

[121] Nahdia Majeed, Maria Saladina, Michal Krompiec, Steve Greedy, Carsten Deibel, and Roderick CI MacKenzie. Using deep machine learning to understand the physical performance bottlenecks in novel thin-film solar cells. *Advanced Functional Materials*, 30(7):1907259, 2020.

[122] Bart Olsthoorn, R Matthias Geilhufe, Stanislav S Borysov, and Alexander V Balatsky. Band gap prediction for large organic crystal structures with machine learning. *Advanced Quantum Technologies*, 2(7-8):1900023, 2019.

[123] Fan Li, Xiaoqi Peng, Zuo Wang, Yi Zhou, Yuxia Wu, Minlin Jiang, and Min Xu. Machine learning (ml)-assisted design and fabrication for solar cells. *Energy & Environmental Materials*, 2(4):280–291, 2019.

[124] Arbia Riahi Sfar, Zied Chtourou, and Yacine Challal. A systemic and cognitive vision for IoT security: A case study of military live simulation and security challenges. In *2017 International Conference on Smart, Monitored and Controlled Cities (SM2C)*, pages 101–105, 2017.

[125] J. Cabra, D. Castro, J. Colorado, D. Mendez, and L. Trujillo. An iot approach for wireless sensor networks applied to e-health environmental monitoring. In *2017 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*, pages 578–583, 2017.

[126] Nabih Alaoui, Jean-Pierre Cances, and Vahid Meghdadi. Energy consumption in wireless sensor networks for network coding structure and ARQ protocol. In *2015 International Conference on Electrical and Information Technologies (ICEIT)*, pages 317–321, 2015.

[127] Francesco Tonolini and Fadel Adib. Networking across boundaries: Enabling wireless communication through the water-air interface. In *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication*, SIGCOMM '18, page 117–131, New York, NY, USA, 2018. Association for Computing Machinery.

[128] Nacer Khalil, Mohamed Riduan Abid, Driss Benhaddou, and Michael Gerndt. Wireless sensors networks for internet of things. In *2014 IEEE Ninth International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP)*, pages 1–6, 2014.

[129] Kofi Sarpong Adu-Manu, Nadir Adam, Cristiano Tapparello, Hoda Ayatollahi, and Wendi Heinzelman. Energy-harvesting wireless sensor networks (eh-wsns): A review. *ACM Trans. Sen. Netw.*, 14(2), apr 2018.

[130] Sonam Lata, Shabana Mehfuz, and Shabana Urooj. Secure and reliable wsn for internet of things: Challenges and enabling technologies. *IEEE Access*, 9:161103–161128, 2021.

[131] Hashem Elsaraf, Mohsin Jamil, and Bishwajeet Pandey. Techno-economic design of a combined heat and power microgrid for a remote community in newfoundland canada. *IEEE Access*, 9:91548–91563, 2021.

[132] Hyun-Il Lim. A linear regression approach to modeling software characteristics for classifying similar software. In *2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)*, volume 1, pages 942–943, 2019.

[133] Daniel Mavilo Calderon Nieto, Erik Alex Papa Quiroz, and Miguel Angel Cano Lengua. A systematic literature review on support vector machines applied to regression. In *2021 IEEE Sciences and Humanities International Research Conference (SHIRCON)*, pages 1–4, 2021.

[134] James E Payne. A survey of the electricity consumption-growth literature. *Applied energy*, 87(3):723–731, 2010.

[135] Aishwarya S Mundada, Kunal K Shah, and Joshua M Pearce. Levelized cost of electricity for solar photovoltaic, battery and cogen hybrid systems. *Renewable and Sustainable Energy Reviews*, 57:692–703, 2016.

[136] Wei Shen, Xi Chen, Jing Qiu, Jennifier A Hayward, Saad Sayeef, Peter Osman, Ke Meng, and Zhao Yang Dong. A comprehensive review of variable renewable energy levelized cost of electricity. *Renewable and Sustainable Energy Reviews*, 133:110301, 2020.

[137] Vivien Kizilcec, Catalina Spataru, Aldo Lipani, and Priti Parikh. Forecasting solar home system customers' electricity usage with a 3d convolutional neural network to improve energy access. *Energies*, 15(3):857, 2022.

[138] Oliver Schmidt, Adam Hawkes, Ajay Gambhir, and Iain Staffell. The future cost of electrical energy storage based on experience rates. *Nature Energy*, 2(8):1–8, 2017.

[139] VY Kondaiah, B Saravanan, P Sanjeevikumar, and Baseem Khan. A review on short-term load forecasting models for micro-grid application. *The Journal of Engineering*, 2022.

[140] Seth B Darling, Fengqi You, Thomas Veselka, and Alfonso Velosa. Assumptions and the levelized cost of energy for photovoltaics. *Energy & environmental science*, 4(9):3133–3139, 2011.

[141] Amir Mosavi and Abdullah Bahmani. Energy consumption prediction using machine learning; a review. 2019.

[142] David W Bian, Sterling M Watson, Natasha C Wright, Sahil R Shah, Tonio Buonassisi, Devarajan Ramanujan, Ian M Peters, et al. Optimization and design of a low-cost, village-scale, photovoltaic-powered, electrodialysis reversal desalination system for rural india. *Desalination*, 452:265–278, 2019.

[143] Waqas Khan, Shalika Walker, and Wim Zeiler. Improved solar photovoltaic energy generation forecast using deep learning-based ensemble stacking approach. *Energy*, 240:122812, 2022.

[144] Gokhan Mert Yagli, Dazhi Yang, and Dipti Srinivasan. Automatic hourly solar forecasting using machine learning models. *Renewable and Sustainable Energy Reviews*, 105:487–498, 2019.

[145] Varaha Satra Bharath Kurukuru, Ahteshamul Haque, Mohammed Ali Khan, Subham Sahoo, Azra Malik, and Frede Blaabjerg. A review on artificial intelligence applications for grid-connected solar photovoltaic systems. *Energies*, 14(15):4690, 2021.

[146] Borut Del Fabbro, Aljoša Valentinčič, and Andrej F Gubina. An adequate required rate of return for grid-connected pv systems. *Solar Energy*, 132:73–83, 2016.

[147] Jacqueline Yujia Tao and Anton Finenko. Moving beyond lcoe: impact of various financing methods on pv profitability for sids. *Energy Policy*, 98:749–758, 2016.

[148] Sergio Shimura, Rafael Herrero, Marcelo Knorich Zuffo, and Jose Aquiles Baesso Grimoni. Production costs estimation in photovoltaic power plants using reliability. *Solar Energy*, 133:294–304, 2016.

[149] Nathaniel Heck, Courtney Smith, and Eric Hittinger. A monte carlo approach to integrating uncertainty into the levelized cost of electricity. *The Electricity Journal*, 29(3):21–30, 2016.

[150] Benjamin Pillot, Sandro de Siqueira, and João Batista Dias. Grid parity analysis of distributed pv generation using monte carlo approach: The brazilian case. *Renewable Energy*, 127:974–988, 2018.

[151] Jose-Ramon Rodriguez-Ossorio, Alberto Gonzalez-Martinez, Miguel de Simon-Martin, Ana-Maria Diez-Suarez, Antonio Colmenar-Santos, and Enrique Rosales-Asensio. Levelized cost of electricity for the deployment of solar photovoltaic plants: The region of león (spain) as case study. *Energy Reports*, 7:199–203, 2021.

[152] DL Talavera, P Pérez-Higueras, JA Ruíz-Arias, and EF Fernández. Levelised cost of electricity in high concentrated photovoltaic grid connected systems: spatial analysis of spain. *Applied energy*, 151:49–59, 2015.

[153] Manasseh Obi, Shauna M Jensen, Jennifer B Ferris, and Robert B Bass. Calculation of levelized costs of electricity for various electrical energy storage systems. *Renewable and Sustainable Energy Reviews*, 67:908–920, 2017.

[154] Niko Lukač, Sebastijan Seme, Katarina Dežan, Borut Žalik, and Gorazd Štumberger. Economic and environmental assessment of rooftops regarding suitability for photovoltaic systems installation based on remote sensing data. *Energy*, 107:854–865, 2016.

[155] Tom M Mitchell and Tom M Mitchell. *Machine learning*, volume 1. McGraw-hill New York, 1997.

[156] Thomas G Dietterich. Machine learning. *Annual review of computer science*, 4(1):255–306, 1990.

[157] Yong Yu, Xiaosheng Si, Changhua Hu, and Jianxun Zhang. A review of recurrent neural networks: Lstm cells and network architectures. *Neural computation*, 31(7):1235–1270, 2019.

[158] Kenneth Gilbert. An arima supply chain model. *Management Science*, 51(2):305–310, 2005.

[159] Mi Dong, Ya Li, Dongran Song, Jian Yang, Mei Su, Xiaofei Deng, Lingxiang Huang, MH Elkholy, and Young Hoon Joo. Uncertainty and global sensitivity analysis of levelized cost of energy in wind power generation. *Energy Conversion and Management*, 229:113781, 2021.

[160] EIA. Environmental investigation agency, 2022. https://eia-international.org/ [Accessed: (10 December 2022)].

[161] IRENA. International renewable energy agency, 2022. https://irena.org/ [Accessed: (10 December 2022)].

[162] U.S. Department of Economic Analysis. Bureau of economic analysis, 2022. https://www.bea.gov/ [Accessed: (10 December 2022)].

[163] IEA. International energy agency, 2022. https://iea.org/ [Accessed: (10 December 2022)].

[164] Fuat Egelioglu, AA Mohamad, and H Guven. Economic variables and electricity consumption in northern cyprus. *Energy*, 26(4):355–362, 2001.

[165] Yong Soo Kim. Comparison of the decision tree, artificial neural network, and linear regression methods based on the number and types of independent variables and sample size. *Expert Systems with Applications*, 34(2):1227–1234, 2008.

[166] Michael Thompson. Regression methods in the comparison of accuracy. *Analyst*, 107(1279):1169–1180, 1982.

[167] Jeffrey T Dellosa, Marcellin Jay C Panes, and Randell U Espina. Techno-economic analysis of a 5 mwp solar photovoltaic system in the philippines. In *2021 IEEE International Conference on Environment and Electrical Engineering and 2021 IEEE Industrial and Commercial Power Systems Europe (EEEIC/I&CPS Europe)*, pages 1–6. IEEE, 2021.

[168] Seul-Gi Kim, Jae-Yoon Jung, and Min Kyu Sim. A two-step approach to solar power generation prediction based on weather data using machine learning. *Sustainability*, 11(5):1501, 2019.

[169] Amy JC Trappey, Paul PJ Chen, Charles V Trappey, and Lin Ma. A machine learning approach for solar power technology review and patent evolution analysis. *Applied Sciences*, 9(7):1478, 2019.

[170] Benjamin Sanchez-Lengeling and Alán Aspuru-Guzik. Inverse molecular design using machine learning: Generative models for matter engineering. *Science*, 361(6400):360–365, 2018.

[171] Changyeon Lee, Seungjin Lee, Geon-U Kim, Wonho Lee, and Bumjoon J Kim. Recent advances, design guidelines, and prospects of all-polymer solar cells. *Chemical reviews*, 119(13):8028–8086, 2019.

[172] Ming-Gang Ju, Min Chen, Yuanyuan Zhou, Jun Dai, Liang Ma, Nitin P Padture, and Xiao Cheng Zeng. Toward eco-friendly and stable perovskite materials for photovoltaics. *Joule*, 2(7):1231–1241, 2018.

[173] Emmanuel Gbenga Dada, Joseph Stephen Bassi, Haruna Chiroma, Adebayo Olusola Adetunmbi, Opeyemi Emmanuel Ajibuwa, et al. Machine learning for email spam filtering: review, approaches and open research problems. *Heliyon*, 5(6):e01802, 2019.

[174] Timo Huuhtanen and Alexander Jung. Predictive maintenance of photovoltaic panels via deep learning. In *2018 IEEE Data Science Workshop (DSW)*, pages 66–70. IEEE, 2018.

[175] Chuanlang Zhan and Jiannian Yao. More than conformational "twisting" or "coplanarity": molecular strategies for designing high-efficiency nonfullerene organic solar cells. *Chemistry of Materials*, 28(7):1948–1964, 2016.

[176] Xiaoshun Zhang, Shengnan Li, Tingyi He, Bo Yang, Tao Yu, Haofei Li, Lin Jiang, and Liming Sun. Memetic reinforcement learning based maximum power point tracking design for pv systems under partial shading condition. *Energy*, 174:1079–1090, 2019.

[177] Nicola Gasparini, Alberto Salleo, Iain McCulloch, and Derya Baran. The role of the third component in ternary organic solar cells. *Nature Reviews Materials*, 4(4):229–242, 2019.

[178] Livia Faes, Siegfried K Wagner, Dun Jack Fu, Xiaoxuan Liu, Edward Korot, Joseph R Ledsam, Trevor Back, Reena Chopra, Nikolas Pontikos, Christoph Kern, et al. Automated deep learning design for medical image classification by health-care professionals with no coding experience: a feasibility study. *The Lancet Digital Health*, 1(5):e232–e242, 2019.