**REST Journal on Data Analytics and Artificial Intelligence**

**Vol: 3(1), March 2024**

**REST Publisher; ISSN: 2583-5564 (Online)**

**Website: https://restpublisher.com/journals/jdaai/**

**DOI: https://doi.org/10.46632/jdaai/3/1/10**

# Comparative Performance Analysis of Solar Energy Systems using Linear Regression and Support Vector Regression

***R. Saravanan, G. Rathnasamy, M. Mohamed Farook, S. Nakul**

*Arasu Polytechnic College, Kumbakonam, Thanjavur District, Tamil Nadu, India.*

*Corresponding Author Email: saravananrajen90@gmail.com

***Abstract:*** *The necessity for sustainability in infrastructure has resulted in the rapid adoption of solar energy systems in building designs. The study aimed to create predictive modeling for the output energy generated in solar energy systems through the application of machine learning models, namely Linear Regression and Support Vector Regression, based on the given dataset with 20 parameters and corresponding parameters such as solar irradiance, panel efficiency, panel area, and ambient temperature, among others. The results from the descriptive statistical analysis showed that the average solar irradiance was 5.65 kWh/m²/day, while the average panel efficiency was 19.28%. The energy output was averaged to be 27.33 kWh/day, while nearly perfect product correlation was established for each input variable and energy output, as shown by $r = 0.98$ and $r = 0.99$ for each corresponding value. The Linear Regression model had a high $R^2$ value, which was 0.9948 for the training set and 0.9767 for the test set, thus proving the efficiency of the model in generalizing the data with minimal over-fitting. The Support Vector Regression model had an extremely high $R^2$ value, which was 0.9973 for the training set, and a high $R^2$ value, which was 0.9505 for the test set, thus validating the efficiency of this model as well. It can be noted that both models had extremely high accuracy in predicting the energy output based on the panel area. Additionally, it can be noted that both models had extremely low error values for all parameters, namely RMSE, MAE, and MSLE. This further validates the fact that machine learning models like LR and SVR are highly efficient in the prediction of solar energy output. This research can be said to be contributing to the overall objective of incorporating data science techniques into the design of Building Integrated Photovoltaic (BIPV) systems.*

***Keywords:*** *Solar energy systems; Building-integrated photovoltaic (BIPV); Linear regression; Support Vector Regression; Energy output forecasting; Machine learning.*

## 1. INTRODUCTION

Solar-integrated buildings, which can also be termed as solar-powered or solar-enabled buildings, can be seen as a major step forward in building design and management. The emergence of solar-integrated buildings can be seen as a response to challenges and issues, including climate change, energy consumption, and the need to move away from traditional sources of energy [1]. In this case, buildings can be seen as playing a critical role, especially with regards to energy consumption and environmental sustainability. In this case, solar energy is seen as one of the most available and used alternatives to traditional sources of energy, especially in the field of building constructions. The idea of solar-integrated design goes beyond the mere installation of solar panels on the rooftops of buildings. Instead, it is seen as part of the design of the buildings, reaching a point where energy generation becomes part of the physical structure of the building. These kinds of structures can be used to generate energy from the sun using photovoltaic cells and a combination of different sources of energy [3]. The major focus of this initiative is to build structures that are "energy-efficient," "eco-friendly," and "self-sufficient." This initiative is focused on raising structures to a level that can be termed "net-zero" or "energy-positive," i.e., structures that can produce more energy than they consume [4]. An important breakthrough in this area of research is the development of Building-Integrated Photovoltaic (BIPV) technology. This technology differs from the conventional solar panel technology, which is installed *on top* of the structure. This technology integrates the photovoltaic technology directly and seamlessly into the structure of the building. In this regard, the structure of the building comprises roofs, walls, windows, skylights, and shading screens [6]. This technology is useful in minimizing the number of materials used in building constructions and can make it easier to integrate solar panels into building designs. Especially in urban areas where there is limited space, this technology can act as an alternative for architects and engineers to ensure that buildings have both beauty and efficiency. The other advantages of solar technology, especially in relation to architectural designs, is the development of luminescent solar concentrators. The device is composed of a glass surface that is transparent or semi-transparent. The advantage of this type of solar concentrator is that it is capable of absorbing solar energy and channeling it to photovoltaic cells that are located on its edges. The advantage of this type of solar panel is that it can produce electricity without interfering with the amount of natural light that is able to get into the building. The device is very useful in high-rise buildings that have large glass surfaces [9]. Significant advancements have been made to the properties of quantum dots, which have greatly enhanced the ability of this form of solar concentrator to absorb light. However, there have been some challenges to the development of this form of solar concentrator, which include durability, energy loss, and manufacturing. To enhance the visual compatibility of solar concentrators, various forms of photovoltaic modules have been developed. It is evident that conventional solar panels have a negative visual impact. This is mainly due to the dark and opaque nature of the conventional solar panel. This has led to the development of various forms of semi-transparent and colored solar panels. These types of solar panels are designed to allow the passage of light and to enhance the visual compatibility of the solar panel [12]. Despite the fact that the efficiency of these solar panels is lower, the visual flexibility of these solar panels has increased the opportunities for the integration of solar systems into the context of modern architecture [13]. To increase the visual compatibility of these solar panels, researchers have also introduced semi-transparent and colored solar panels. Conventional solar panels are considered visually incompatible due to their dark colors. To a certain extent, the dark colors of these solar panels have reduced the opportunities for the integration of solar panels into the context of modern architecture. However, the introduction of semi-transparent and colored solar panels has reduced the above-mentioned problem, as these solar panels can be used as a facade, windows, or as a decorative item of the building. Although the efficiency of these solar panels is lower, the opportunities for the integration of these solar panels into the context of modern architecture have increased. The utilization of solar energy in buildings has increased far beyond the generation of electricity [14]. Solar thermal technology is increasingly being used to satisfy the demands of thermal energy. In the context of buildings, especially residential, public, and institutional buildings, which have high demands for thermal energy, solar thermal technology assumes great importance. Improvements have been made in the solar thermal technology used, especially with respect to the material used, coatings, and storage. These improvements have made the service life of solar thermal technology longer. In regions with high sunlight, solar thermal technology assumes great importance, especially when the demand for electricity is high. In such regions, the role of solar thermal technology has been highly relevant with respect to the development of sustainable development [15].

## 2. METHODOLOGY

In the analysis of solar energy systems, the parameters include solar radiation, efficiency of the solar panel, and the area of the solar panel, ambient temperature, and energy. Solar radiation refers to the amount of sunlight available. The efficiency of the solar panel refers to the ability of the solar panel to convert sunlight into electrical energy. The area of the solar panel refers to the total surface area of the solar panel, which determines the amount of solar radiation received. Ambient temperature refers to the temperature of the environment, which affects the efficiency of the solar panel. Energy refers to the electrical output of the system.

**Linear Regression** One of the most popular techniques used in the field of biomedical predictive modeling is linear regression. The reason behind the popularity of this technique is the simplicity and ease of understanding it offers. Linear regression is a technique in which a linear relationship is established between the independent and dependent variables. Therefore, linear regression is a powerful technique in the field of exploratory data analysis and basic modeling. The simplicity and ease of understanding of linear regression have made this technique a popular choice in the field of clinical and regulatory decision-making. However, the linear relationship that is a basic requirement of linear regression makes this technique unsuitable for use in the field of biomedical predictive modeling, as complex physiological processes cannot be modeled using a linear relationship. Linear regression is a technique that is adversely affected by the presence of outliers, and the impact of outliers is significant in the case of biological data. However, despite these limitations, linear regression still plays an important role as a baseline method for evaluating the performance of more sophisticated modeling approaches. By moving from simple linear regression to multiple linear regression, it is possible to simultaneously model the effects of multiple variables such as temperature, pH levels, mechanical stress, and physiological parameters. This makes it easier to model the complex nature of biomedical systems.

**Support Vector Regression** as developed by Wapnick et al. in the 1990s, has become a mainstay in the field of machine learning research and application. These models have seen extensive application in various fields, including biometrics, economic modeling, and bioinformatics. The field of bioinformatics has seen a surge in the application of SVM models, largely due to the exponential growth in available datasets. The recent interest in using SVM models in chemical classification problems as well as calibration procedures has seen much excitement in the analytical chemistry research community. There is a growing body of literature on the comparison of SVM models with traditional models in chemistry. Application Scope and Data Manipulation: The application of SVM models is mostly seen with limited variables, as is common in analytical chemistry problems. The application is not limited to such constraints. Using appropriate data pre-processing techniques, complex datasets with large dimensions can be effectively handled by SVMs. The versatility of this algorithm makes it more appropriate to handle complex analytical problems. The regression-based variant of SVM is called support vector regression (SVR), which is based on Vapnik's statistical learning theory. SVR is a supervised learning technique that has received considerable recognition due to its outstanding performance in various application domains, such as document classification, computer vision, genetics, financial risk assessment, etc. The main advantage of SVR is that it is based on a solid mathematical foundation, is less prone to local optima, and can effectively handle large datasets with growing sample size as the dimensionality increases.

## 3. ANALYSIS AND DISSECTION

**TABLE 1.** Solar Energy Systems

| | Solar Irradiance | Panel Efficiency | Panel Area | Ambient Temperature | Energy |
|---|---|---|---|---|---|
| count | 20.0000 | 20.0000 | 20.0000 | 20.0000 | 20.0000 |
| mean | 5.6500 | 19.2800 | 30.5500 | 33.2500 | 27.3300 |
| std | 0.7082 | 0.6810 | 3.9666 | 2.6730 | 4.8278 |
| min | 4.5000 | 18.0000 | 24.0000 | 29.0000 | 19.6000 |
| 25% | 5.0750 | 18.8750 | 27.7500 | 31.0000 | 23.7750 |
| 50% | 5.7500 | 19.3500 | 30.5000 | 33.5000 | 27.7000 |
| 75% | 6.2250 | 19.8250 | 34.0000 | 35.0000 | 31.6000 |
| max | 6.7000 | 20.3000 | 37.0000 | 38.0000 | 34.5000 |

Table 1 shows that, based on 20 studies, the performance of the solar energy system can be described as consistent. This is because the average solar irradiance of 5.65 kWh/m²/day has low variability, signifying good solar resource availability. In addition, the efficiency of the solar panels is consistent at 19.28%, while the average area of the panels is 30.55 m². Moreover, the ambient temperature is at a moderately high level of 33.25 °C. Finally, the average energy output of 27.33 kWh/day varies depending on irradiance, area, and efficiency.

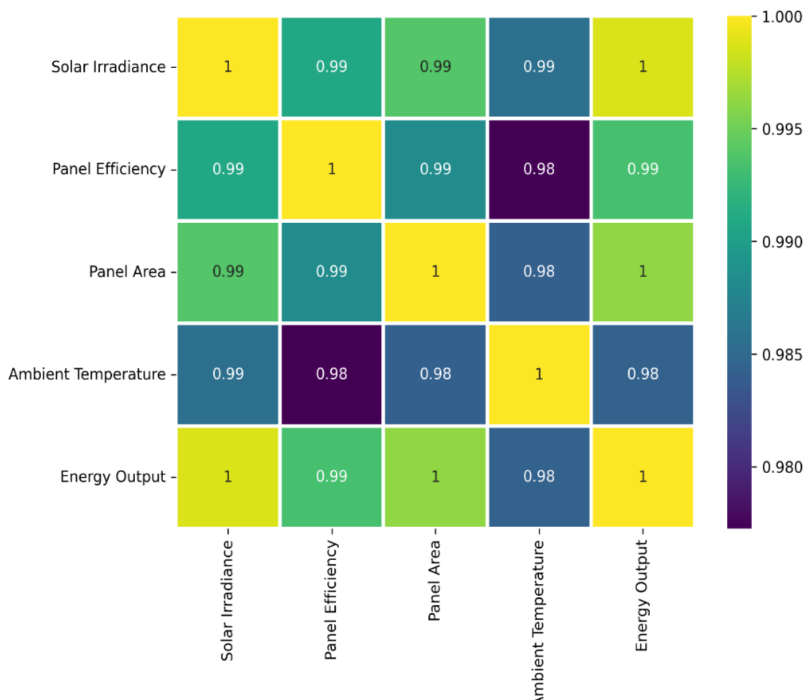


**FIGURE 1.** Solar Energy Systems Correlation heat map

Figure 1 demonstrates that there is a correlation between the scatter matrix, solar irradiance, panel efficiency, panel area, ambient temperature, and energy output. It is evident from the figure that as solar irradiance and panel area increase, energy output is significantly affected by these two factors. Moreover, it is evident from the figure that ambient temperature is also positively correlated;

this is possibly because of high levels of irradiance. It is evident from the histograms that data is uniformly distributed and lacks skewness.

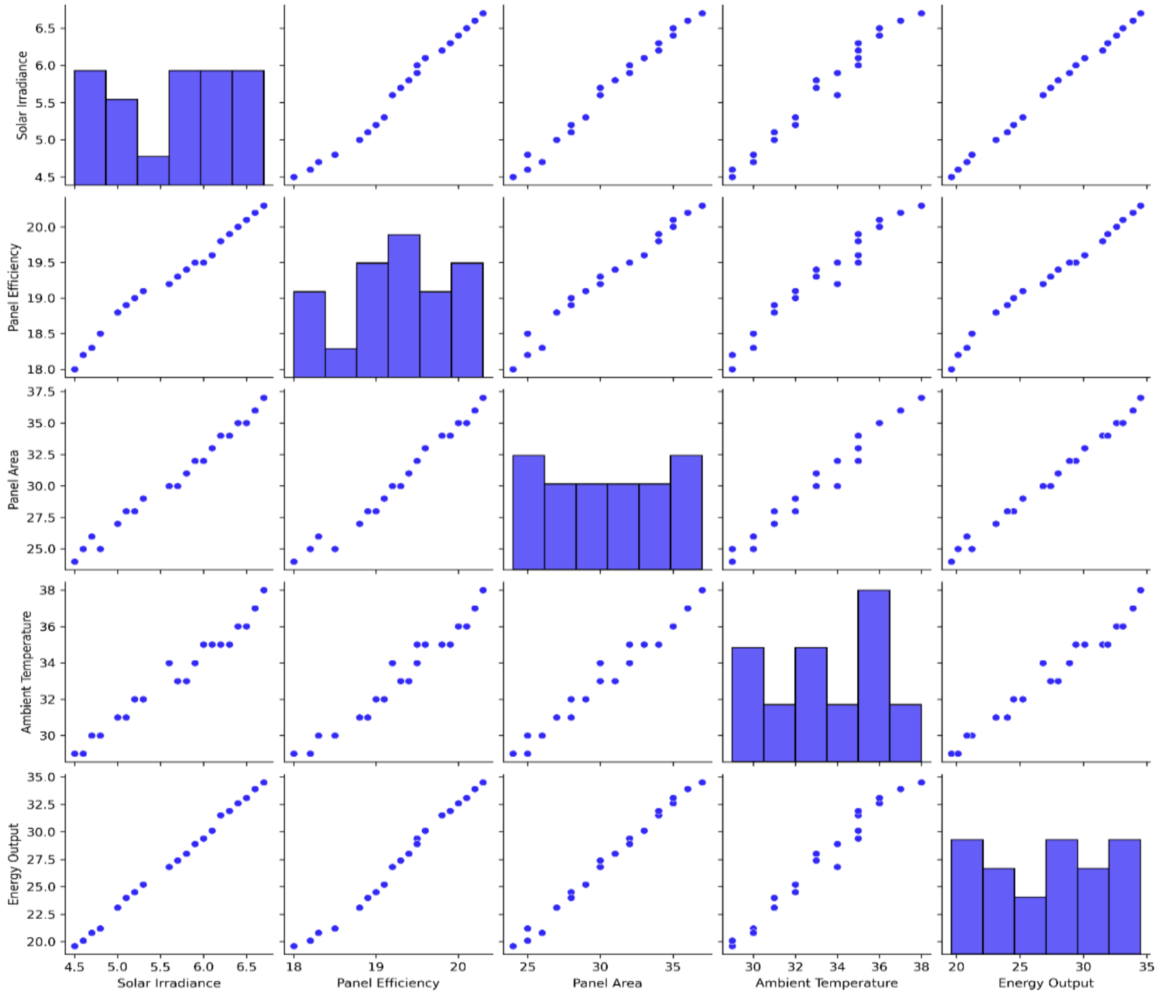


**FIGURE 2.** Solar Energy Systems Effect of Process Parameters

Figure 2 indicates that there are correlations between the scatter matrix, solar irradiance, panel efficiency, panel area, ambient temperature, and energy output. With the increase in irradiance, efficiency, and panel area, there is a significant increase in energy output. Thus, the impact of irradiance, efficiency, and area on energy output can be seen as significant. There is a positive correlation between ambient temperature and energy output. This could be due to the increase in irradiance levels. From the histograms, the distribution of the data is uniform with no significant skewness.

**TABLE 2.** Linear Regression Models Panel Area Train and Test performance metrics

| Linear Regression | Train | Test |
|---|---|---|
| R2 | 0.9948 | 0.9767 |
| EVS | 0.9948 | 0.9825 |
| MSE | 0.0654 | 0.2905 |
| RMSE | 0.2558 | 0.5390 |
| MAE | 0.2026 | 0.4041 |
| Max Error | 0.4337 | 1.2859 |
| MSLE | 0.0001 | 0.0004 |
| Med AE | 0.1733 | 0.2787 |

Table 3 Linear Regression demonstrates superior predictive performance, exhibiting the highest $R^2$ and EVS values across both the training and testing datasets. Low error metrics indicate accurate predictions and robust generalization. The minimal discrepancy between the training and testing results suggests negligible over fitting.

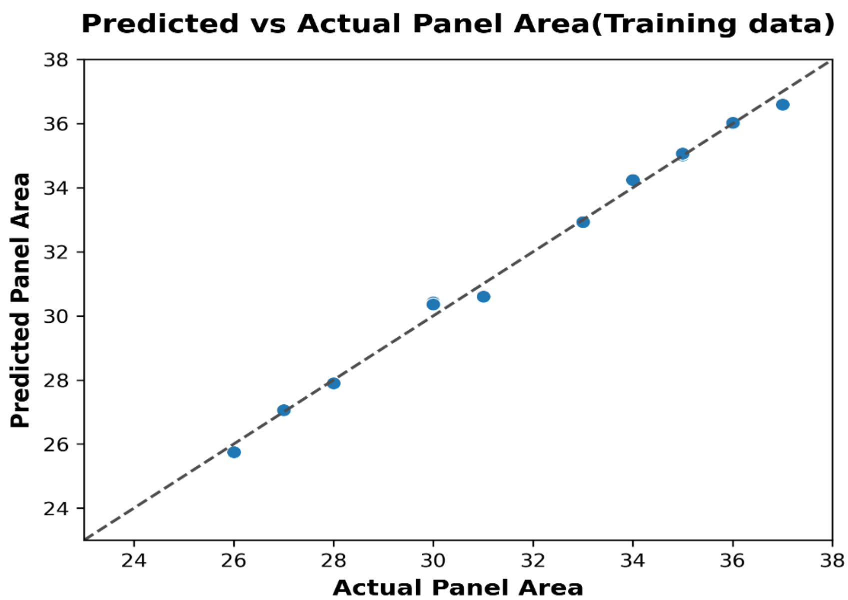


**FIGURE 3.** Linear Regression Panel Area Training

Figure 3 the scatter plot in figure 3 demonstrates how closely the predicted panel area matches the actual values in the training data set. Most of the data points in the plot lie close to the diagonal reference line, signifying high predictive accuracy and strong linear correlation. Slight variations from this line indicate low error, signifying how well the data set is being predicted.

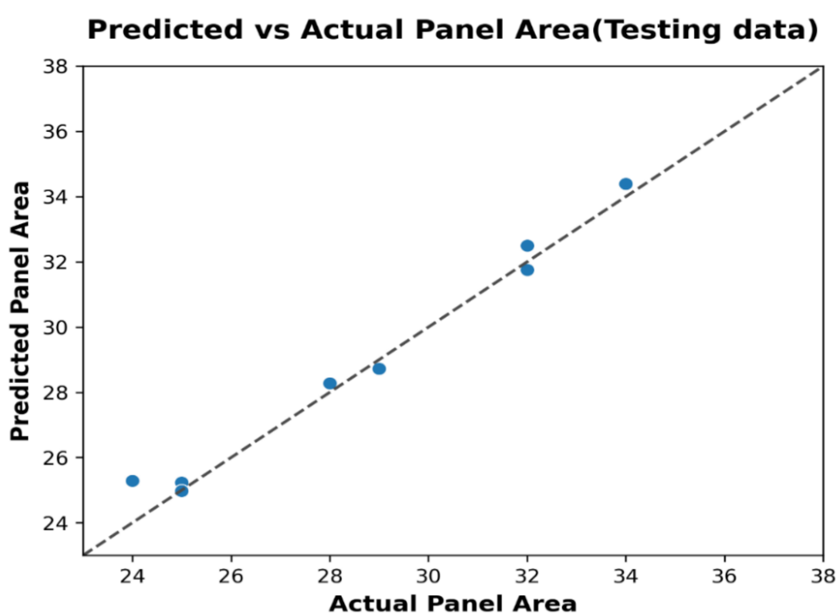


**FIGURE 4.** Linear Regression Panel Area Testing

Figure 4, the test scatter plot, indicates that the values obtained for the panel area prediction are close to the actual values. Most of the values are close to the diagonal line, showing that the prediction is accurate and the model is able to generalize well. There is a slight variation in the values, but the error is minimal. This indicates that the model is able to work well and is stable.

**TABLE 3**. Support Vector Regression Models Panel Area Train and Test performance metrics

| Support Vector Regression | Train | Test |
|---|---|---|
| R2 | 0.9973 | 0.9505 |
| EVS | 0.9974 | 0.9627 |
| MSE | 0.0336 | 0.6179 |
| RMSE | 0.1834 | 0.7861 |
| MAE | 0.1543 | 0.6187 |
| Max Error | 0.4285 | 1.5616 |
| MSLE | 0.0000 | 0.0009 |
| Med AE | 0.1005 | 0.3834 |

As presented in Table 3, it is clear that the Support Vector Regression (SVR) model has higher predictive performance compared to other models for the solar energy dataset. The performance on the training set shows that the model is very accurate, as indicated by a very high value for the coefficient of determination ($R^2 = 0.9973$), implying that almost all the variance is explained by the model. The same is confirmed by the explained variance score, where EVS is equal to 0.9974, implying that there is a very good fit and low dispersion from the actual values. The error metrics, such as MSE, RMSE, and MAE, are very low, implying that there is high accuracy and precise learning. The performance on the test set is very good, as shown by a value of $R^2$ equal to 0.9505 and EVS equal to 0.9627. The slightly lower values compared to those obtained on the training set indicate that there is good performance. The moderate increase in error metrics implies that there is a low degree of over fitting. The very low value for MSLE and median absolute error confirms consistency.

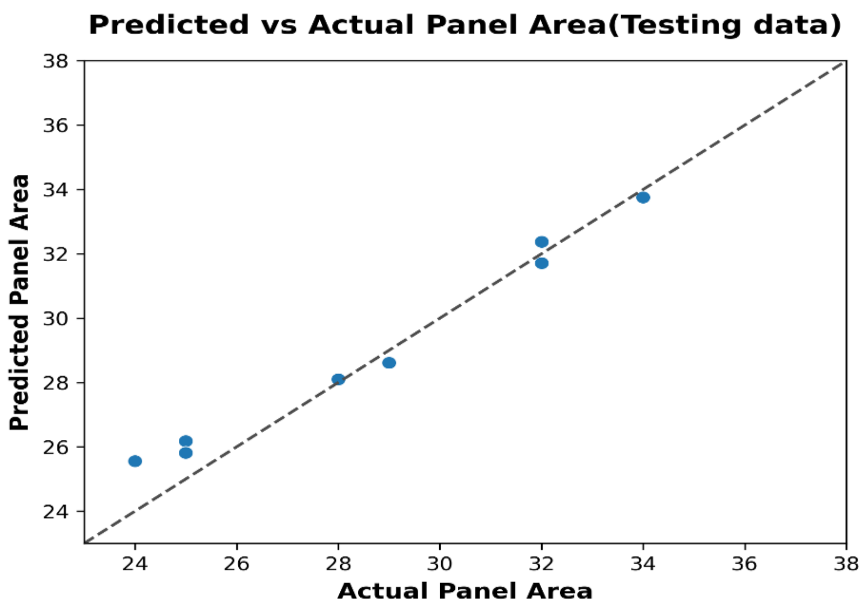


**FIGURE 5**. Support Vector Regression Panel Area Training

Figure 5, the plot for the prediction and actual panel area for the test dataset, indicates a high level of agreement between the model prediction and the actual values. All the data points are clustered tightly around the ideal line, showing that the model is a good predictor of the panel area for the data in the testing set. A slight variation from the ideal line indicates a prediction error, especially for smaller values of panel area, where the model over predicts. However, it is evident that the prediction error is small and does not indicate any form of bias. A tight cluster around the ideal line indicates a high level of model generalization and robustness.

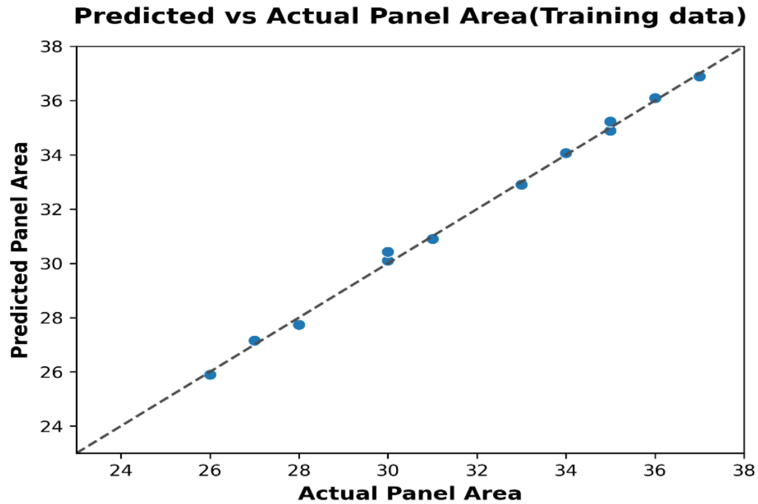


**FIGURE 6.** Decision Tree Regression Circuit Depth Testing

In Figure 6, which is a plot of predicted and actual values of panel areas for the training data set, there is an excellent fit between the predicted and actual values. It is noted that nearly all the points lie on or near the reference line, which indicates high accuracy of the model's predictions for all values of panel areas. The low degree of scatter around the diagonal line indicates very low prediction errors, and it is noted that the model has learned the values effectively. It is also noted that there is no indication of overestimation or underestimation of values, which indicates a balanced performance of the model. Additionally, the lack of outlier points indicates a stable performance of the regression model during training. This is also indicated by high training performance values, such as an almost perfect value of $R^2$.

## 4. CONCLUSION

The applicability of machine learning techniques, namely Linear Regression and Support Vector Regression, was successfully demonstrated for the prediction of the output of the integrated solar energy systems. This was achieved by a thorough analysis of the parameters affecting the output of the solar energy systems. Highly accurate predictions were obtained for the output of the solar energy systems, thereby validating the applicability of the machine learning techniques. Moreover, descriptive statistical analysis showed that solar radiation, panel efficiency, and panel area are strongly correlated to energy output since their correlation coefficients are almost equal to 1.00. The almost perfect multicollinearity between input variables further emphasizes the significance of feature selection and validation in designing predictive models for solar energy systems. In addition, a correlation heat map further supported the finding that, in the dataset, solar radiation and panel area are the most significant factors in energy production. The Linear Regression model obtained $R^2$ values of 0.9948 on the training set and 0.9767 on the testing set, thus ensuring proper generalization with minimum over fitting. The low error rates of the LR model, with RMSE of 0.2558 on the training set and 0.5390 on the testing set, thus prove the robustness of the model, making it highly applicable in the assessment of solar energy performance. The Support Vector Regression (SVR) model obtained an $R^2$ of 0.9973 on the training set, whereas the testing set obtained an $R^2$ of 0.9505. Thus, the model ensured robust performance, as indicated by the minimum difference between the performance of the model during the training and testing phases. However, the LR model showed better consistency, thus making it more applicable. Both models were able to accurately predict the values of the panel area during the training and testing phases. This was confirmed by the data points clustering closely to the 45-degree reference line. This therefore validated the potential of machine learning to support the design, optimization, and monitoring of Building Integrated Photovoltaic (BIPV) systems.

## REFERENCES

[1]. Richards, Bryce S., and Ian A. Howard. "Luminescent solar concentrators for building integrated photovoltaics: opportunities and challenges." Energy & Environmental Science 16, no. 8 (2023): 3214-3239.

[2]. Hussain, Shibna, Santosh Kumar Sharma, and Shiv Lal. "Feasible synergy between hybrid solar PV and wind system for energy supply of a green building in Kota (India): A case study using iHOGA." Energy Conversion and Management 315 (2024): 118783.

[3]. Xie, Mengju, Minghao Wang, Hua Zhong, Xinxin Li, Baofeng Li, Thushini Mendis, and Shen Xu. "The impact of urban morphology on the building energy consumption and solar energy generation potential of university dormitory blocks." Sustainable Cities and Society 96 (2023): 104644.
[4]. Omeiza, Lukman Ahmed, Muhammad Abid, Anitha Dhanasekaran, Yathavan Subramanian, Veena Raj, Kateryna Kozak, Ukashat Mamudu, and Abul Kalam Azad. "Application of solar thermal collectors for energy consumption in public buildings–An updated technical review." Journal of Engineering Research 12, no. 4 (2024): 994-1010.
[5]. Li, Jin, Tinglei Yue, and Yin Zhang. "Experimental and simulation study on the modified shape coefficient of building integrated thermal considerations for heating buildings in solar-enriched areas." Solar Energy 264 (2023): 112060.
[6]. Ashrafi, Negar, Rouhollah Ahmadi, and Alireza Zahedi. "Technical, economical, and environmental scenario-based modeling of the building equipped with ground source heat pump (GSHP) and solar system." Energy and Buildings 289 (2023): 113048.
[7]. Chandrika, V. S., Alagar Karthick, Nallapaneni Manoj Kumar, P. Manoj Kumar, B. Stalin, and M. Ravichandran. "Experimental analysis of solar concrete collector for residential buildings." International Journal of Green Energy 18, no. 6 (2021): 615-623.
[8]. Vidhya Prasanth, Chandrasekar Raja, M. Ramachandran, Sathishkumar Mani, "Advances in Water Purification Technologies Innovations and Challenges Using TOPSIS Method", Building Materials and Engineering Structures, Vol 4(1) 2026, 25–37.
[9]. Li, Zhenpeng, Tao Ma, Hongxing Yang, Lin Lu, and Ruzhu Wang. "Transparent and colored solar photovoltaics for building integration." Solar RRL 5, no. 3 (2021): 2000614.
[10]. Muhammed, Eslam, Adel El-Shazly, and Salem Morsy. "Building rooftop extraction using machine learning algorithms for solar photovoltaic potential estimation." Sustainability 15, no. 14 (2023): 11004.
[11]. Thanh, Tung Nguyen, Phap Vu Minh, Kien Duong Trung, and Tuan Do Anh. "Study on performance of rooftop solar power generation combined with battery storage at office building in northeast region, Vietnam." Sustainability 13, no. 19 (2021): 11093.
[12]. Smaisim, Ghassan F., Azher M. Abed, Salema K. Hadrawi, and Ali Shamel. "Modeling and thermodynamic analysis of solar collector cogeneration for residential building energy supply." Journal of Engineering 2022, no. 1 (2022): 6280334.
[13]. Paudel, Bhola, Niraj Regmi, Parlad Phuyal, Deependra Neupane, M. Imtiaz Hussain, Dae Hyun Kim, and Sagar Kafle. "Techno-economic and environmental assessment of utilizing campus building rooftops for solar PV power generation." International Journal of Green Energy 18, no. 14 (2021): 1469-1481.
[14]. Huang, Yaoyao, Zhangyu Yuan, Jia Yang, Shungao Yin, Aihui Liang, Gang Xie, Chuizheng Feng et al. "Highly efficient perovskite solar cells by building 2D/3D perovskite heterojuction in situ for interfacial passivation and energy level adjustment." Science China Chemistry 66, no. 2 (2023): 449-458.
[15]. Chisale, Sylvester William, and Patrick Mangani. "Energy audit and feasibility of solar PV energy system: Case of a commercial building." Journal of Energy 2021, no. 1 (2021): 5544664.
[16]. Yang, Jie, Kit-Ying Chan, Harun Venkatesan, Eunyoung Kim, Miracle Hope Adegun, Jeng-Hun Lee, Xi Shen, and Jang-Kyo Kim. "Superinsulating BNNS/PVA composite aerogels with high solar reflectance for energy-efficient buildings." Nano-micro letters 14, no. 1 (2022): 54.
[17]. Lunt, Mark. "Introduction to statistical modelling: linear regression." Rheumatology 54, no. 7 (2015): 1137-1140.
[18]. Buckley, Jonathan, and Ian James. "Linear regression with censored data." Biometrika 66, no. 3 (1979): 429-436.
[19]. Mateos, Gonzalo, Juan Andrés Bazerque, and Georgios B. Giannakis. "Distributed sparse linear regression." IEEE Transactions on Signal Processing 58, no. 10 (2010): 5262-5276.
[20]. Groeneboom, Piet, and Kim Hendrickx. "Current status linear regression." The Annals of Statistics 46, no. 4 (2018): 1415-1444.
[21]. Chen, Chun-Houh, and Ker-Chau Li. "Can SIR be as popular as multiple linear regression?" Statistica Sinica (1998): 289-316.
[22]. Vidhya Prasanth, Chitra Periyasamy, M. Ramachandran, Libiya Saravanan, "Biomedical Applications of Polymer-Composite Materials using Fuzzy ARAS Method" Journal on Materials and its Characterization, 4(4), 2025, 4-14.
[23]. Brereton, Richard G., and Gavin R. Lloyd. "Support vector machines for classification and regression." Analyst 135, no. 2 (2010): 230-267.
[24]. Ceperic, Ervin, Vladimir Ceperic, and Adrijan Baric. "A strategy for short-term load forecasting by support vector regression machines." IEEE Transactions on Power Systems 28, no. 4 (2013): 4356-4364.
[25]. Shamshirband, Shahaboddin, Dalibor Petković, Hossein Javidnia, and Abdullah Gani. "Sensor data fusion by support vector regression methodology—a comparative study." IEEE Sensors Journal 15, no. 2 (2014): 850-854.
[26]. Funt, Brian, and Weihua Xiong. "Estimating illumination chromaticity via support vector regression." (2004).
[27]. Musicant, David R., and Alexander Feinberg. "Active set support vector regression." IEEE Transactions on Neural Networks 15, no. 2 (2004): 268-275.
[28]. Ustun, B. Support vector machines: facilitating the interpretation and application. Sl: sn, 2009.
[29]. Grimm, Michael, Kristian Kroschel, and Shrikanth Narayanan. "Support vector regression for automatic recognition of spontaneous emotions in speech." In 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07, vol. 4, pp. IV-1085. IEEE, 2007

 