RSC Sustainability

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *RSC Sustainability*, 2024, **2**, 3520

Identifying the best ML model for predicting the bandgap in a perovskite solar cell†

Nita Samantaray, *[a] Arjun Singh *[a] and Anu Tonk[b]

Perovskite solar cells (PSCs) have gained attention for their characteristics of high efficiency and commercial viability. However, the efficiency of a PSC depends on various factors. One such important parameter is the bandgap of the active layer as it plays an important role in PSCs with regards to the amount of light absorption. Thus, it influences the overall performance of the solar cell. It is important to predict the bandgap of the active layer in PSCs to achieve an effective fabrication process. In this study, we compared six machine learning (ML) models to predict the bandgap. The models were created using a dataset of 500 devices, such as $MAPbI_3$, $FAPbI_3$, $CsSnI_3$ and $CsMAPbI_3$, obtained from The Perovskite Database Project. These models were further validated using a different dataset of 50 devices. The models were created using ML methods: random forest, gradient boosting regressor, k-nearest neighbours (KNN), AdaBoost, Gaussian process regressor, and bagging. The feature parameters considered for the models were the A coefficient, B coefficient, and C coefficient, out of various other parameters such as the perovskite dimension, perovskite thickness, perovskite deposition temperature, and perovskite deposition time. The random forest model showed better results compared to other models with a low mean absolute error (MAE) of 0.000775, low mean squared error (MSE) of 0.00000920, and high coefficient of determination (r^2) of 0.9994.

Received 10th July 2024
Accepted 4th October 2024

DOI: 10.1039/d4su00370e

rsc.li/rscsus

Sustainability spotlight

Our research aims to enhance the efficiency of perovskite solar cells (PSCs) by accurately predicting the bandgap of the active layer—a critical factor in light absorption and overall performance. We evaluated six machine learning models, using a dataset of 500 devices, to predict the bandgap. Among these, the random forest model demonstrated superior performance with a low mean absolute error (MAE) of 0.000775, a low mean squared error (MSE) of 0.00000920, and a high coefficient of determination (r^2) of 0.9994. Our work supports the UN Sustainable Development Goals: affordable and clean energy (SDG 7); industry, innovation, and infrastructure (SDG 9); and climate action (SDG 13).

1. Introduction

PSCs have emerged as a promising material in the field of photovoltaics owing to their exceptional efficiency and potential for commercial viability.[1,2] These solar cells utilize the unique properties of perovskite materials to convert sunlight into electricity with remarkable efficiency. An important factor influencing the performance of PSCs is the bandgap of the active layer, which decides the amount of light absorption and thus the overall efficiency of the solar cell. An accurate prediction of the bandgap is essential for optimizing the fabrication process and enhancing the performance of PSCs.[3]

To improve the efficiency of PSCs, extensive research into understanding and optimizing the properties of perovskite materials has been carried out. The bandgap of the active layer is a key parameter that directly impacts the absorption spectrum and photon-to-electron conversion efficiency of PSCs.[4] Consequently, predicting and optimizing the bandgap is essential for achieving high-performance PSCs. Traditional methods for predicting the bandgap rely on complex theoretical models and experimental techniques, which can be time-consuming and resource-intensive.[5] In recent years, ML techniques have emerged as powerful tools for predicting material properties with high accuracy and efficiency.[6]

In 2024, Miah *et al.* emphasized the critical role of bandgap tuning in enhancing both the performance and stability of PSCs, offering insights into mechanisms that optimize efficiency while addressing degradation factors. Perovskite materials exhibit excellent optoelectronic properties and efficiency in solar cells, but their low stability hinders commercialization.[7]

In 2024, Ghosh *et al.* published a study that focused on predicting the bandgaps of nitride perovskites using four

[a]*Department of Applied Sciences, The Northcap University, Gurugram, India. E-mail: nita20asd006@ncuindia.edu; arjunsingh@ncuindia.edu*

[b]*Department of Multidisciplinary Engineering, The Northcap University, Gurugram, India*

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d4su00370e

machine learning (ML) models: multi-layer perceptron (MLP), gradient boosted decision tree (GBDT), support vector regression (SVR), and random forest regression (RFR). The models were trained on 1563 nitride perovskites with bandgaps between 1.0 and 3.1 eV.[8]

Sadhu *et al.* recently suggested that ML models accurately forecast PSC parameters, with key features such as the grain size, band gap, and electron/hole mobility driving performance optimization for commercialization. The study focused on the analysis and prediction of the performance of PSCs using machine learning techniques.[9]

In our study, we investigated the accuracy of various ML models in predicting the bandgap of perovskite materials used in PSCs. We utilized a dataset from The Perovskite Database[3] comprising information on 500 devices, including different perovskite compositions such as $MAPbI_3$, $FAPbI_3$, $CsSnI_3$, and $CsMAPbI_3$. The dataset contains a range of feature parameters relevant to the fabrication process, including perovskite dimensions, coefficients, thickness, deposition temperature, and deposition time. Our primary objective is to compare the performance of six different ML models in predicting the bandgap of the perovskite material, and identify the most accurate and reliable predictive model.

To achieve our objective, we employed six ML methods: random forest, gradient boosting regressor, k-nearest neighbors (KNN), AdaBoost, Gaussian process regressor, and bagging. These models are trained on a dataset of 500 devices, and subsequently validated using a separate dataset consisting of 50 devices. The performance of each model is evaluated based on key metrics, such as the mean absolute error (MAE), mean squared error (MSE), and coefficient of determination (r^2). Additionally, we analysed the feature importance of the selected parameters to gain insights into their influence on bandgap prediction.

Our results demonstrate that the random forest model outperforms the other ML models in predicting the bandgap of perovskite materials for PSCs, exhibiting low MAE and MSE values, and a high coefficient of determination (r^2). This highlights the accuracy of the random forest model in capturing the complex relationships between the input parameters and the bandgap of the active layer. Furthermore, our analysis sheds light on the importance of specific feature parameters in determining the bandgap, providing valuable guidance for optimizing the fabrication process of PSCs. Overall, this study contributes to advancing the understanding and predictive capabilities of ML techniques in the field of perovskite solar cells, paving the way for enhanced device performance and widespread adoption of this promising renewable energy technology.

The random forest regression model in Fig. 1 accurately predicts the bandgap, enabling researchers to systematically explore and fine-tune perovskite compositions, dimensions, and deposition parameters.

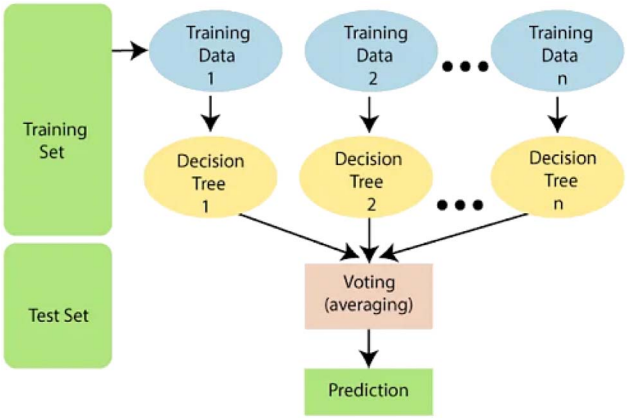

Fig. 1 Diagram illustrating the learning processes of the random forest model.

2. Methodology

2.1 Data collection and preprocessing

Data were taken from The Perovskite Database comprising 500 perovskite solar cells (PSCs), including various compositions, such as $MAPbI_3$, $FAPbI_3$, $CsSnI_3$, and $CsMAPbI_3$. This dataset has various feature parameters relevant to the fabrication process, such as perovskite dimensions, coefficients (A, B, C), thickness, deposition temperature, and deposition time. Each device in the dataset was characterized by its corresponding bandgap, giving the target variable for our predictive models. To ensure the relatability of the dataset, data preprocessing steps were performed, including handling missing values, removing outliers, and normalizing feature scales.[10] Some of the prominent PSCs and its actual bandgap on the collected dataset are mentioned in Table 1 for reference.

Heatmap analysis was used to identify the feature parameters in predicting the bandgap of perovskite materials for perovskite solar cells (PSCs). Heatmaps provide a visual representation of the correlation between each input feature parameter and the target variable (bandgap).[15] The original dataset for this study includes essential parameters, such as the power conversion efficiency (PCE), open-circuit voltage, short-circuit current, and fill factor. However, since the bandgap pertains specifically to the active layer, *i.e.*, the perovskite material, rather than the entire perovskite solar cell, we focused on parameters directly associated with the active layer.[16] These parameters include the perovskite dimension, perovskite thickness, perovskite deposition temperature, perovskite deposition time, and the A, B, and C coefficients.

Further, by analysing the heatmap as mentioned in Fig. 2, we were able to identify the feature parameters with the strongest correlations to the bandgap, indicating their importance in the predictive model. Hence, we identified that the A coefficient, B

Table 1 Type of perovskites (ABC_3) with the A, B and C coefficients and their actual bandgap

Sl. No.	A	B	C	Perovskites	Actual bandgap	References
1	MA	Pb	I	$MAPbI_3$	1.6	11
2	FA	Pb	I	$FAPbI_3$	1.6	12
3	Cs	Sn	I	$CsSnI_3$	1.3	13
4	Cs; MA	Pb	I	$CsMAPbI_3$	1.5	14

coefficient, and C coefficient have higher and positive heatmap coefficients, *i.e.*, 0.70, 0.83, 0.63, respectively, and are strongly correlated with the bandgap compared to the other parameters. Hence, we considered the A coefficient, B coefficient, and C coefficient for training the models and performing further analysis.

Visualizing the feature importance through a heatmap provided significant clarity on how each parameter influenced the bandgap prediction. The heatmap allowed for an intuitive understanding of the relative significance of each feature by displaying their importance in a clear, color-coded matrix. This method made it easier to identify the most impactful factors and how they correlated with the predicted bandgap.

The heatmap provided a visual and analytical tool that enhanced the interpretability of feature importance, aiding researchers in focusing on the most important parameters for improving bandgap prediction and the overall performance of PSCs.

2.2 Model training and evaluation

We selected six ML models known for their robustness and effectiveness in regression tasks: random forest, gradient boosting regressor, k-nearest neighbors (KNN), AdaBoost, Gaussian process regressor, and bagging. These models were chosen based on their ability to capture the complexity between input features and target variables, which is essential for accurately predicting the bandgap of perovskite materials in PSCs. Each model was trained on the pre-processed dataset of 500 devices, and tested on a separate dataset having 50 devices.

The performance and accuracy of the models were then verified using the following metrics.

2.2.1 Coefficient of determination (r^2). r^2 measures the correlation between different sets of variables, and indicates what it takes for the model to fit the data. It ranges from 0 to 1, in which 1 signifies a perfect fit and high reliability, while 0 suggests weak correlation and poor generalization. Eqn (1) for r^2 compares the squared variations between the predicted and target outputs to the squared differences between the target outputs and their mean.[11]

$$r^2 = 1 - \frac{\left(Y_j - Y_i^p\right)^2}{\left(Y_j - \underline{Y_j}\right)^2} \quad (1)$$

where Y_j accounts for the values which are targeted, and Y_i^p denotes the predicted outputs that we get from the model. The variable Y_j denotes the mean of Y_j.

2.2.2 Root mean square error (RMSE). RMSE quantifies the average of difference between the predicted and target outputs. It provides an overall accuracy of the generalized regression model. Eqn (2) is computed by taking the square root of the

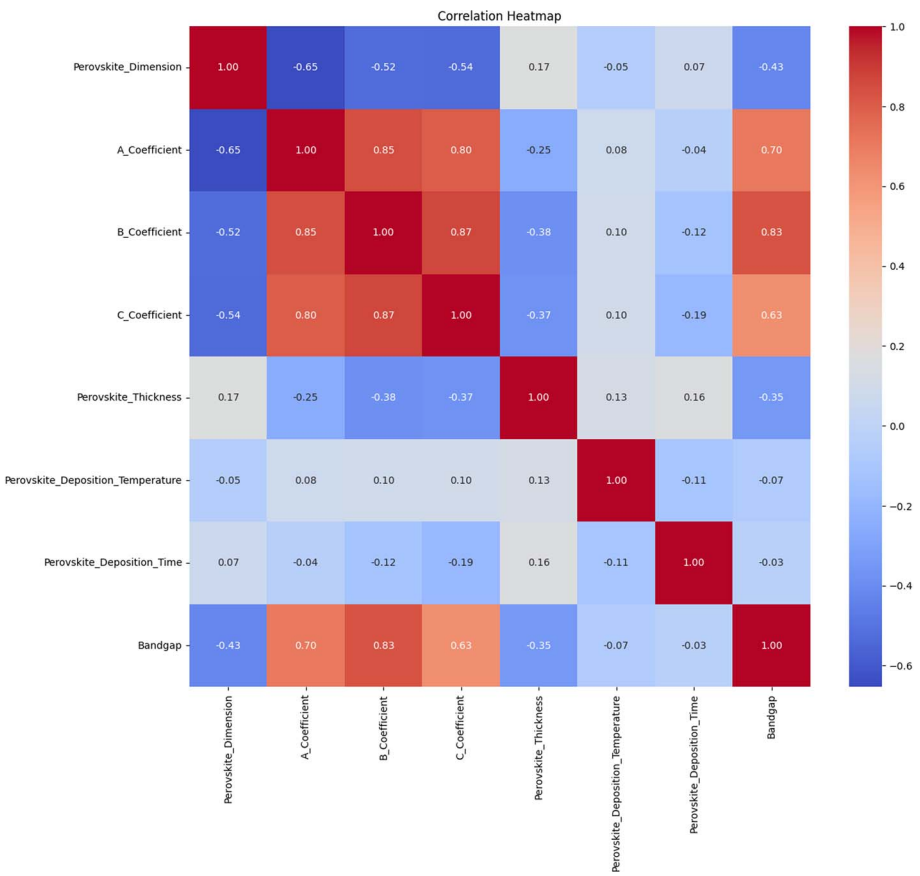

Fig. 2 . Heatmap analysis to determine the feature parameters.

Table 2 Comparison of the experimental data of different ML models

Sl. No.	ML model	Mean absolute error (MAE)	Mean squared error (MSE)	Coefficient of determination (r^2)
1.	Random forest	0.000775	0.00000920	0.9994
2.	Gradient boosting regressor	0.0222	0.0041	0.8841
3.	k-Nearest neighbors (KNN)	0.0365	0.0143	0.5914
4.	AdaBoost	0.0283	0.0029	0.9163
5.	Gaussian process regressor	0.0351	0.0082	0.7662
6.	Bagging	0.0448	0.0162	0.6673

average value of the squared differences between the predicted and target outputs.[12]

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\left(Y_j - Y_j^p\right)^2} \quad (2)$$

2.2.3 Mean absolute error (MAE). The mean absolute error calculates the average absolute difference between the predicted and target outputs for continuous variables. It represents the average errors without considering their direction. The calculation for MAE is shown in eqn (3), and is obtained by taking the average of the differences between the predicted and target outputs.[13]

$$\mathrm{MAE} = \frac{1}{n}\left|Y_j - Y_j^p\right| \quad (3)$$

These metrics offer quantitative assessments of a model's generalization capacity, providing insights into its correlation with the data, accuracy in predicting target outputs, and overall error magnitude. By analysing these metrics, researchers and practitioners can evaluate and compare the efficiency of different trained models.

3. Results and discussion

Among the various machine learning ML models (such as random forest, gradient boosting regressor, k-nearest neighbours (KNN), AdaBoost, Gaussian process regressor, and bagging) evaluated in this study, random forest emerged as the most effective ML model for the prediction of the bandgap of PSCs. A comparative analysis shown in Table 2 below revealed that random forest exhibited higher accuracy and precision, due its low mean absolute error (MAE) of 0.000775, low mean squared error (MSE) of 0.00000920 and high coefficient of determination (r^2) of 0.9994, indicating its exceptional ability to explain the variance in the target variable. The study was performed using perovskite solar cell parameters, such as the A coefficient, B coefficient and C coefficient.

In this analysis of machine learning models for predicting the bandgap of perovskite materials in PSCs, we visualized the performance metrics of each model using a line graph with markers. The above table offers a clear and concise comparison of the mean absolute error (MAE), mean squared error (MSE), and coefficient of determination (r^2) across different machine learning models. Each line in the graph corresponds to a specific performance metric, while the markers indicate the values associated with individual machine learning models. It reveals distinct patterns in model performance, with random forest demonstrating the lowest MAE and MSE values, as well as the highest r^2 coefficient among all models evaluated.

Table 2 enhances the comprehensibility of our findings, enabling researchers and practitioners in the field to easily discern the relative performance of different machine learning approaches for predicting the bandgap of perovskite materials in PSCs.

4. Conclusion

This study analysed the effectiveness of ML models in predicting the bandgap of perovskite materials for use with PSCs. By comparing six different ML models, including random forest, gradient boosting regressor, k-nearest neighbours (KNN), AdaBoost, Gaussian process regressor, and bagging, random forest was found to be more effective. The performance of random forest was highly effective, with a low mean absolute error (MAE) of 0.000775, low mean squared error (MSE) of 0.00000920, and high coefficient of determination (r^2) of 0.9994. These results highlight the potential of ML techniques in accurately predicting critical parameters, such as the bandgap of perovskite materials, thereby helping in the development of high-efficiency PSCs.

As future scope, there are several aspects for research in this domain that can be explored. Firstly, research can be directed towards optimizing the feature parameters used in the ML models to further enhance prediction accuracy. Additionally, studies can explore more efficient ML algorithms and ensemble techniques that can even yield better performance in bandgap prediction.[17] A few advanced machine learning models (such as Neural Networks) or Transformer models (such as PolyNC or polyBERT) can be further taken into consideration for this work.[18–20] Furthermore, exploring the same model in different datasets would provide a more comprehensive understanding of the factors influencing the bandgap variation in PSCs. Overall, continued research in this field holds the promise of advancing the development of efficient and commercially viable perovskite solar cells.

This model also can be integrated into the material design workflow. For instance, when new compositions of perovskite materials (*e.g.*, mixed halides or organic-inorganic hybrids) are proposed, the model can quickly estimate the bandgap without

needing extensive experimental trials. This could significantly reduce the time and cost associated with experimental material characterization, allowing for more efficient screening of potential parameters for high-efficiency solar cells.

In addition, the model could be used to guide the fabrication process by offering real-time predictions of bandgap during deposition stages, helping engineers maintain optimal conditions. This would not only improve device performance, but also ensure reproducibility across different manufacturing batches.

Data availability

The data supporting this article are available as part of the ESI† and are available on the journal's homepage in sufficient detail to enable the reproducibility of experiments. The primary dataset comprises 500 perovskite solar cell (PSC) devices, including $MAPbI_3$, $FAPbI_3$, $CsSnI_3$, and $CsMAPbI_3$, whereas the validation dataset consists of 50 PSC devices. These datasets include feature parameters, such as the A coefficient, B coefficient, C coefficient, perovskite dimension, perovskite thickness, perovskite deposition temperature, and perovskite deposition time. Additionally, the machine learning models developed in this study, including random forest, gradient boosting regressor, k-nearest neighbors (KNN), AdaBoost, Gaussian process regressor, and bagging, are available as part of the ESI† and on the journal's homepage in sufficient detail to enable the reproducibility of experiments. We are committed to ensuring transparency and reproducibility of research, and therefore, all ESI† required for this study are made openly accessible.

Conflicts of interest

The authors declare no conflict of interest.

Acknowledgements

Nita Samantaray, Arjun Singh and Anu Tonk contributed equally to this article. The authors thank the NorthCap University, Gurugram, India, for their assistance in writing this article.

References

1 N. S. Kumar and K. C. B. Naidu, A review on perovskite solar cells (PSCs), materials and applications, *J. Materiomics*, 2021, 7(5), 940–956, DOI: 10.1016/j.jmat.2021.04.002.
2 X. Liu, Z. Wu, X. Fu, L. Tang, J. Li, J. Gong and X. Xiao, Highly efficient wide-band-gap perovskite solar cells fabricated by sequential deposition method, *Nano Energy*, 2021, **86**, 106114, DOI: 10.1016/j.nanoen.2021.106114.
3 Perovskite Database, Perovskite Solar Cell Database, https://www.perovskitedatabase.com/.
4 T. Zdanowicz, T. Rodziewicz and M. Zabkowska-Waclawek, *Sol. Energy Mater. Sol. Cells*, 2005, **87**, 757–769.
5 C. Zhu, J. Ni, Z. Yang, Y. Sheng, J. Yang and W. Zhang, *Comput. Theor. Chem.*, 2022, **1217**, 113872.
6 H. He, Y. Wang, Y. Qi, Z. Xu, Y. Li and Y. Wang, *Nano Energy*, 2023, **118**, 108965.
7 Md. H. Miah, M. U. Khandaker, Md. B. Rahman, M. Nur-E-Alam and M. A. Islam, *RSC Adv.*, 2024, **14**, 15876–15906.
8 S. Ghosh and J. Chowdhury, *RSC Adv.*, 2024, **14**, 6385–6397.
9 D. Sadhu, D. Dattatreya, A. Deo, K. Tarafder and D. De, *J. Alloys Compd. Commun.*, 2024, **3**, 100022.
10 V. Vakharia, I. E. Castelli, K. Bhavsar and A. Solanki, *Phys. Lett. A*, 2021, **422**, 127800.
11 A. E. Karrar, *Indones. J. Electr. Eng. Inform.*, 2022, **10**(2), 375–384.
12 M. Caputo, N. Cefarin, A. Radivo, N. Demitri, L. Gigli, J. R. Plaisier, M. Panighel, G. Di Santo, S. Moretti, A. Giglia, M. Polentarutti, F. De Angelis, E. Mosconi, P. Umari, M. Tormen and A. Goldoni, *Sci. Rep.*, 2019, **9**, 15159.
13 C. Tao, Y. Wei, J. Zhang, Y. Cao, S. Wang, L. Xu, K. Wen, J. Wang, Z. Kuang, X. Wang, W. Huang, Q. Peng and J. Wang, *J. Phys. Chem. Lett.*, 2023, **14**, 3805–3810.
14 M. N. Islam, J. Podder and M. L. Ali, *RSC Adv.*, 2021, **11**, 39553–39563.
15 Z. Hui, M. Wang, J. Chen, X. Yin, Y. Yue and J. Lu, *J. Phys.: Condens.Matter*, 2024, **36**, 355901.
16 T. Ibn-Mohammed, S. C. L. Koh, I. M. Reaney, A. Acquaye, G. Schileo, K. B. Mustapha and R. Greenough, *Renewable Sustainable Energy Rev.*, 2017, **80**, 1321–1344.
17 I. H. Sarker, *SN Comput. Sci.*, 2021, **2**, 160.
18 I. O. Oboh, U. H. Offor and N. D. Okon, *Energy Rep.*, 2022, **8**, 973–988.
19 H. Qiu, L. Liu, X. Qiu, X. Dai, X. Ji and Z.-Y. Sun, *Chem. Sci.*, 2024, **15**, 534–544.
20 C. Kuenneth and R. Ramprasad, *Nat. Commun.*, 2023, **14**, 4099.