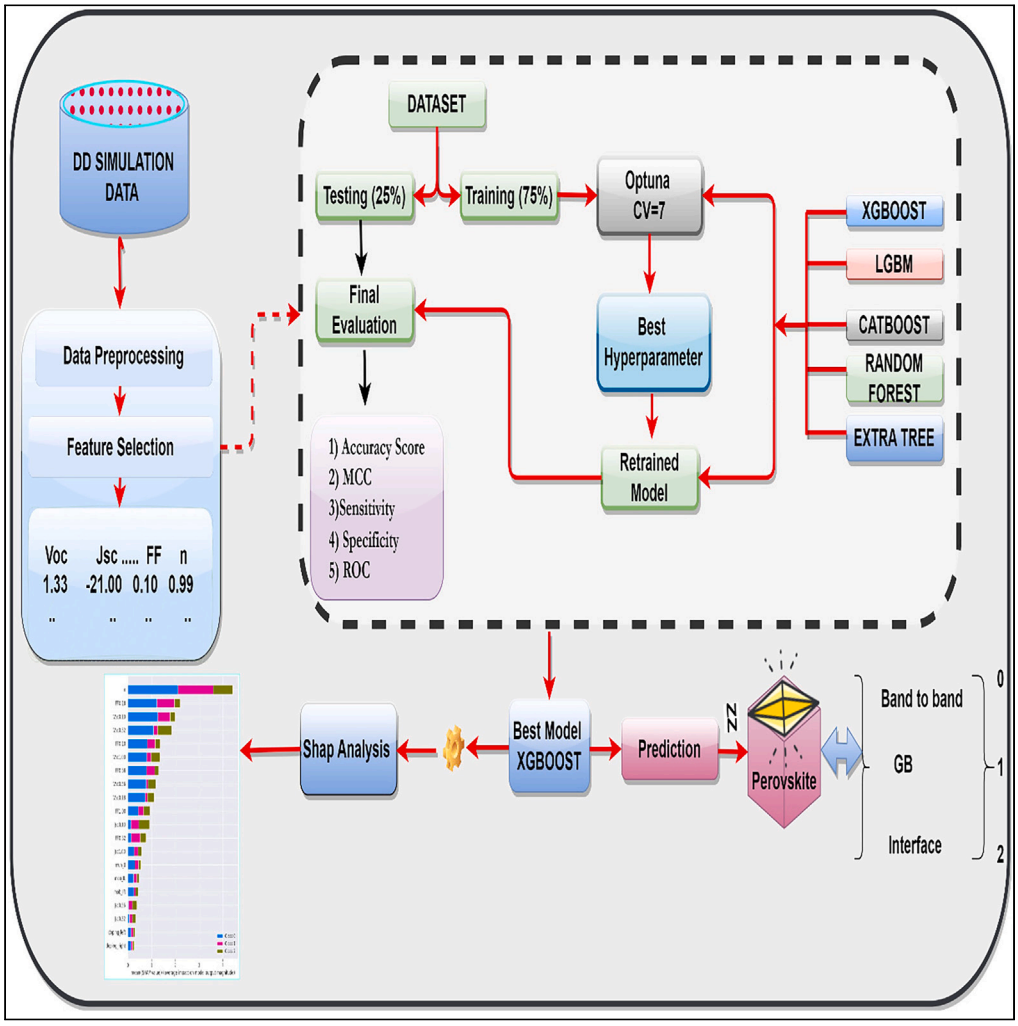
iScience

CellPress
OPEN ACCESS

Article

Unveiling dominant recombination loss in perovskite solar cells with a XGBoost-based machine learning approach



Basir Akbar, Hilal Tayara, Kil To Chong

hilaltayara@jbnu.ac.kr (H.T.)
kitchong@jbnu.ac.kr (K.T.C.)

Highlights

In-depth analysis of dominant recombination losses of perovskite solar cell

Implemented cutting edge techniques of machine learning (ML)

Offer top-tier ML model trained on a large dataset

Reliable experimental validated ML model

Akbar et al., iScience 27, 109200
March 15, 2024
https://doi.org/10.1016/j.isci.2024.109200

iScience

CellPress
OPEN ACCESS

Article

Unveiling dominant recombination loss in perovskite solar cells with a XGBoost-based machine learning approach

Basir Akbar,[1] Hilal Tayara,[2,5,*] and Kil To Chong[3,4,*]

SUMMARY

Remarkable and intelligent perovskite solar cells (PSCs) have attracted substantial attention from researchers and are undergoing rapid advancements in photovoltaic technology. These developments aim to create highly efficient energy devices with fewer dominant recombination losses within the realm of third-generation solar cells. Diverse machine learning (ML) algorithms implemented, addressing dominant losses due to recombination in PSCs, focusing on grain boundaries (GBs), interfaces, and band-to-band recombination. The extreme gradient boosting (XGBoost) classifier effectively predicts the recombination losses. Our model trained with 7-fold cross-validation to ensure generalizability and robustness. Leveraging Optuna and shapley additive explanations (SHAP) for hyperparameter optimization and investigate the influence of features on target variables, achieved 85% accuracy on over 2 million simulated data, respectively. Because of the input parameters (light intensity and open-circuit voltage), the performance evaluation measures for the dominant losses caused by the recombination predicted by proposed model were superior to those of state-of-the-art models.

INTRODUCTION

Inorganic and organic perovskites are considered emerging candidates and have attracted significant attention as alternatives to conventional solar cell components owing to their superior properties and remarkable performance as light absorption/charge transport layers for the development of photovoltaic technology.[1] Perovskites solar cells (PSCs) have recently undergone significant developments to enhance their power conversion efficiency and lower their production costs compared with conventional solar cells.[2] Efforts are being made to build machine learning (ML) tools for accurately predicting dominant recombination and analyzing PSCs data.[3] ML has gained considerable attention in various sectors, such as health, physics, finance, and transportation, to learn from data and enhance decision-making abilities.[4–7] ML algorithms and artificial intelligence (AI) can be used to reduce the production time frame and explore and discover novel material structures and their physical properties for the efficient use of solar energy.[8,9] Additionally, AI tools or ML algorithms can open new horizons to reveal efficient and innovative techniques for identifying the compositions of various material structures to achieve maximum efficiency.[10] Hence, ML has successfully proven its ability to identify the composition and characteristics of different materials, which is valuable for exploring promising material structures and maximizing the PSCs efficiency.[11,12] Additionally, ML tools can provide a new pathway for researchers to efficiently identify the behavior of material data and predict the characteristics of novel explored materials.[13–15] Salah et al. used ML to reveal electron transport layer (ETL) doping's critical influence on PSC efficiency across various dataset complexities.[16] Nagaswa et al. proposed a random forest (RF)-based ML screening method using a dataset of known polymer characteristics to categorize and identify the performance and applicability of conjugated polymers for solar cell applications.[17] Lu et al. utilized ML to forecast perovskite solar cell performance from experimental inputs, facilitating the creation of highly efficient cell.[18]

Additionally, the key phenomenon in PSCs that affects efficiency is recombination. In 2017, Sherkar et al. improved the stability and efficiency of PSCs by focusing on the significance of grain boundaries (GBs), interface traps, and ions in the recombination process.[19] Wolff et al. recognized the essence of interfaces in the recombination process using energy-level alignments, charge transfer kinetics, and interfacial defects to develop stable and efficient PSCs.[20] However, the open-circuit voltage (V_{oc}) is an important parameter for reducing the dominant losses caused by recombination at the PSC interface. Juan-Pablo et al. demonstrated that PSCs with enhanced V_{oc} minimize the losses caused by recombination at the interface.[21] Additionally, Guo et al. proposed highly efficient PSCs with superior reliability and high V_{oc} by exploring various parameters, including material quality, device layout, and interface engineering.[22] Furthermore, doping significantly influences PSCs

[1]Graduate School of Integrated Energy-AI, Jeonbuk National University, Jeonju 54896, South Korea
[2]School of International Engineering and Science, Jeonbuk National University, Jeonju 54896, South Korea
[3]Advances Electronics and Information Research Center, Jeonbuk National University, Jeonju 54896, South Korea
[4]Department of Electronics and Information Engineering, Jeonbuk National University, Jeonju 54896, South Korea
[5]Lead contact
*Correspondence: hilaltayara@jbnu.ac.kr (H.T.), kitchong@jbnu.ac.kr (K.T.C.)
https://doi.org/10.1016/j.isci.2024.109200

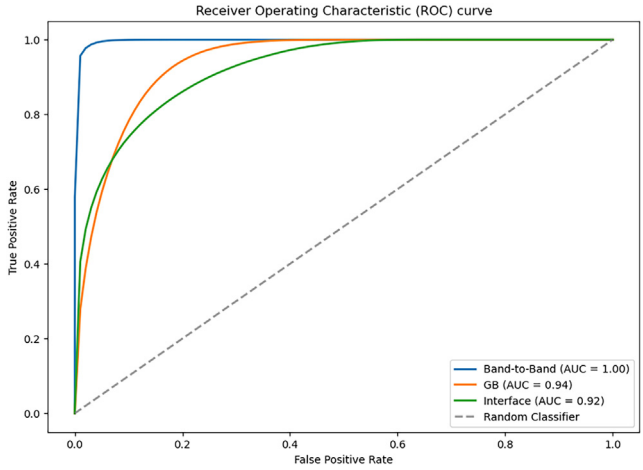


Figure 1. The mean AUC of the proposed model for band-to-band, GB, and interface on the test dataset

efficiency, notably reducing recombination rates. Precise doping concentrations are crucial in both transport layers (TLs), the electron transport layer (ETL) and the hole transport layer (HTL).[23,24] Additionally, the ideality factor is the most important factor for finding the rate of recombination in PSCs.[25] Furthermore, Jiangzhao et al. effectively showcased the recombination mechanism as their causative factor and potential resolutions and also demonstrated the crucial role of the ideality factor in nonradioactive recombination.[26] However, it is difficult to conclude the dominant recombination mechanism based on only ideality factor whether it is band-to-band, GB, or interface. Efforts have been directed to minimize the recombination losses in PSCs considering the important factors V_{oc}, light intensities, doping, mobilities and ideality factor, which plays significant role in determining the recombination losses in PSCs.[23] An urgent study is required to design an ML model to predict the material characteristics and dominant losses from recombination in PSCs to produce efficient and long-lasting PSCs.

To predict the dominant loss in PSCs, ML techniques were used to predict the interface, GBs, and band-to-band recombination. Vincent et al. proposed an approach to predict the dominant recombination losses in PSCs using a band-to-band recombination, GBs, and interface considering important factors $V_{oc,}$ light intensity and ideality factor.[27] This study focuses on effectively predicting the dominant loss in band-to-band recombination compared with that in GB and interfaces. To reduce the dominant recombination losses and enhance PSC performance, ML algorithms must be developed that can predict the dominant efficiency loss in material recombination or at solar cell interfaces across various light intensities by analyzing their current voltage properties and accurately identifying the root cause of efficiency loss by evaluating the data.

This study implemented several ML algorithms, including RF, light gradient boosting machine (LightGBM), extreme gradient boosting (XGBoost), CatBoost, neural network, and extra tree classifiers, to address the multi class classification problem. The objective was to efficiently determine the dominant efficiency losses in GB recombination, interface, and band-to-band recombination using light intensity measurements. The proposed algorithms were carefully trained and optimized considering important features, such as the ideality factor, V_{oc}, fill factor (FF), current density (J_{sc}), doping left, doping right, and mobilities. A drift diffusion (DD) simulation was used to create the previous dataset. The aforementioned ML model was trained using 7-fold cross-validation to ensure its generalizability. The XGBoost classifier outperformed the other implemented ML models and achieved an impressive accuracy score of 85% because of the best Optuna hyperparameter optimizer. The performances of the trained ML algorithms were evaluated using a performance evaluation matrix. The proposed models were compared in terms of specificity, sensitivity, Matthews correlation coefficient (MCC), and accuracy. Additionally, the proposed models were compared with previously reported models. Furthermore, our optimal ML model, validated with experimental data available in previously published literatures. The result demonstrated a robust agreement between the model predictions and experimental outputs establishes itself as the more reliable and efficient ML model. Finally, the selected features of the trained models were validated using SHAP analysis to demonstrate the importance of each feature for predicting the target variable. The proposed approach introduces an efficient method for predicting dominant recombination losses in PSCs.

RESULTS AND DISCUSSION

This study used ML models to predict dominant recombination in PSCs. The models included RF, LightGBM, XGBoost, CatBoost, and extra tree classifiers. To evaluate the performance of these models, four metrics were used: sensitivity, specificity, MCC, and accuracy.

We evaluated the proposed models using a 7-fold cross-validation approach. Figure 1 shows the performance of the XGBoost model in terms of area under the curve (AUC) for each label, including "band-to-band," "GB," and "interface." The mean AUC was calculated individually considering all labels. The mean band-to-band AUC was 1.0, indicating that the model accurately distinguished positive and negative samples in this class. This implies that the model distinguished between the target classes; exceptionally well; similarly, the mean AUC for GB was 0.94, indicating that the model was able to effectively distinguish positive and negative samples for the "GB". The mean AUC for the

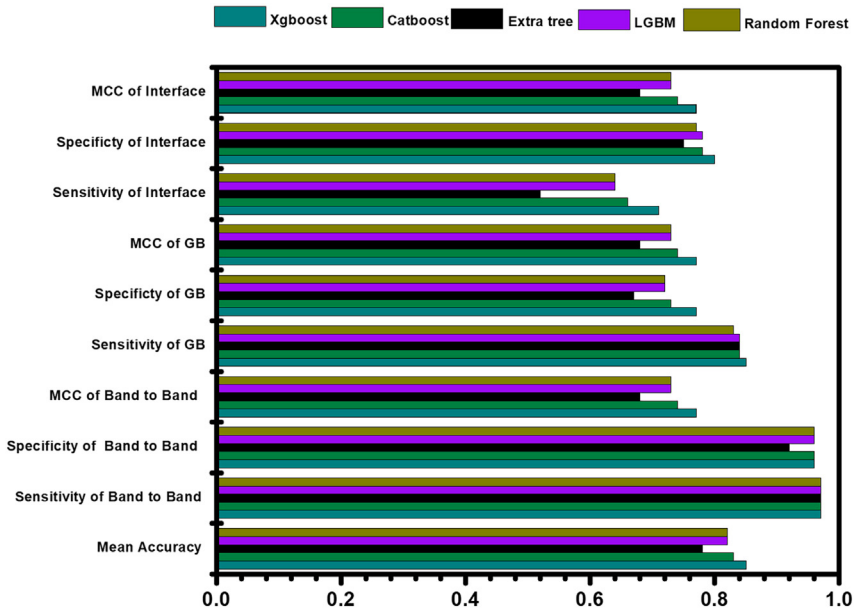


Figure 2. Performance comparison of implemented models: average performance analysis

interface was 0.92. Hence, the proposed model has a strong ability to effectively distinguish classes. Figure S1 (supplemental information) shows the average AUC of the CatBoost classifier for each label, including band-to-band, GB, and interface. Each had an average AUC of 1.0, 0.92, and 0.90, respectively. Figure S2 (supplemental information) shows the average AUC of the extra tree for each label, including the band-to-band, GB, and interface. Each had an average AUC of 0.99, 0.90, and 0.87, respectively. Finally, Figure S3 (supplemental information) shows the average AUC of LGBM, including band-to-band, GB, and interface. Each had an average value of 0.99, 0.92, and 0.91, respectively. To further investigate the performance consistency of the proposed model, the variation in the AUC values was examined. The XGBoost model demonstrated reliable discriminatory power and consistent and robust performance across each fold. Overall, the XGBoost classifier is a viable option for classifying prediction classes because of its previously mentioned stable and consistent behavior.

Based on these findings, XGBoost outperformed all other models used in this study in terms of predicting PSC recombination. Hyperparameter optimization and to find influence of each feature on the target variable Optuna and SHAP were employed to further improve prediction accuracy.

Figure 2 shows the performance of each model, highlighting its specificity, sensitivity, MCC, and accuracy scores after hyperparameter optimization.

Figure 2 illustrates that, XGBoost outperformed all other models, achieving an impressive accuracy score of 85%. The average sensitivity values were notably high for band-to-band (0.97), GB (0.85), and interface (0.71), indicating the model's ability to correctly identify positive instances. Similarly, the model demonstrated excellent band-to-band (0.96), GB (0.77), and interface (0.80) specificity, demonstrating its proficiency in correctly classifying negative instances. Furthermore, the MCC values were consistently favorable for the band-to-band (0.77), GB (0.77), and interface (0.77). Collectively, these findings confirm that the XGBoost classifier outperformed the other models, making it the best choice for this study.

Figure 2 compares the analysis of the proposed ML models. The performance of the developed ML models was evaluated in terms of the mean sensitivity, specificity, MCC, and accuracy score, as shown in Figure 2. Table 1 lists the mean accuracy, sensitivity, specificity, and MCC of each implemented model. However, the band-to-band, GB, and interface labels had mean sensitivities of 0.97, 0.85, and 0.72, respectively, and mean specificities of 0.96, 0.77, and 0.80, respectively. The XGBoost model demonstrated a remarkable mean MCC of 0.77, 0.77, and 0.77, and an accuracy score of 0.85%. The XGBoost model outperformed the RF, LGBM, extra tree, and CatBoost models. According to the previous discussion , the XGBoost model demonstrated superior performance and is a potential candidate for predicting dominant recombination in PSCs.

Additionally, the results and detail of the neural network model was detailed in Table S1 (supplemental information) showing the training and validation accuracies, alongside the corresponding and training and validation losses. Moreover, the Figures S4 and S5 (supplemental information) illustrates the training and validation accuracy curves as well as the training and validation loss curves of the neural network.

Comparison with state-of-the-art studies

This study achieved a remarkable accuracy in predicting the dominant recombination band-to-band ratio with a correctness rate of 0.97. Additionally, the model performed exceptionally well in classifying GB, achieving an accuracy rate of 0.85. Furthermore, it demonstrated

Table 1. Results of the model implemented in this study

		Band to Band			GB			Interface		
Models	Mean Accuracy	MCC	Specificity	Sensitivity	MCC	Specificity	Sensitivity	MCC	Specificity	Sensitivity
XGBoost	0.85	0.77	0.96	0.97	0.77	0.77	0.85	0.77	0.80	0.71
CatBoost	0.83	0.74	0.96	0.97	0.73	0.73	0.84	0.74	0.78	0.66
LightGBM	0.82	0.73	0.96	0.97	0.73	0.72	0.84	0.73	0.78	0.64
Extra tree	0.78	0.68	0.92	0.97	0.68	0.67	0.84	0.68	0.75	0.52
RF	0.82	0.73	0.96	0.97	0.73	0.72	0.83	0.73	0.77	0.64

good predictive capabilities with an interface correctness rate of 0.76. Figure 3 shows a visual representation of the classification report that highlights how well the XGBoost classifier performed. It presents essential evaluation metrics, including precision, recall (sensitivity), and the F1 score, crucial for assessing the performance of an ML model. Precision is employed to assess the model's proficiency in correctly predicting positive outcomes among all instances that it identifies as positive. For band-to-band, GB, and interface, the precision values are 0.97, 0.76, and 0.81, respectively. The F1 score, combining precision and recall, provides a balanced assessment of the model's overall performance. For band-to-band, GB, and interface, the F1 score values are 0.97, 0.81, and 0.76 correspondingly. Table 2 show compares the proposed novel model and the previous approach for predicting the dominant loss in PSC recombination in terms of accuracy and correct prediction across each label. A previous study found that the accuracy of the prediction of dominant losses was not ideal owing to the mixed cases in the GB and interfaces to the overall recombination. Moreover, our study shows a remarkably correct prediction of the GB and band-to-band recombination of the PSCs produced by our proposed model, as shown in Table 2. Figures S6–S8 (supplemental information) show a visual representation of the classification report, which demonstrates the performance of the CatBoost, extra tree, and LGBM classifier.

Metal-halide PSCs have garnered significant attention within the research community, surpassing other alternatives. However, PSCs experience significant non-negligible losses owing to different dominant recombination processes. Dominant recombination losses in PSCs occur mainly at the interfaces and GB. Losses that occur mainly at the interfaces are highly dominant owing to electron-hole pair recombination and reduce PSC efficiency.[19,28] Many studies have been conducted to reduce the dominant losses at the interface and GB. Hence, in this study, we computationally demonstrate that losses at the interface are dominant. However, the proposed model demonstrated peak performance by accurately predicting the dominant recombination losses at the GBs, achieving a correct prediction rate of 0.85. The recombination losses in the GBs can reduce the efficiency and long term stability of the PSCs.[29] Thus, the proposed computational approach yielded superior predictions and provided a sustainable platform for predicting the dominant recombination losses at the GB to fabricate efficient PSCs. Overall, our proposed model outperformed previous experimental approaches and yielded consistent results. Consequently, our approach is a promising and effective method for predicting dominant recombination losses at the band-to-band, GB, and interfaces.

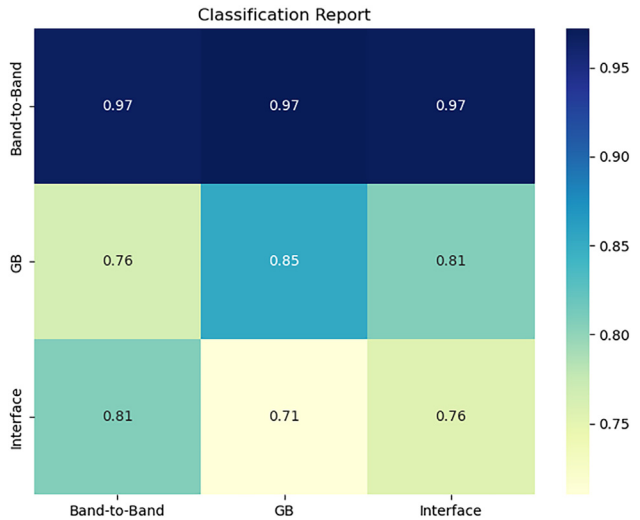


Figure 3. Comprehensive performance analysis: classification report for band-to-band, GB, and interface classes on the test dataset

Table 2. Comparison of the proposed model with previously reported models

Models	Accuracy	Correct prediction band-to-band	Correct prediction GB	Reference
RF	0.82	0.97	0.74	Le Corre et al.[27]
XGBoost	**0.85**	**0.97**	**0.85**	**Our work**

To further validate our analysis, we chose experimental values from previously published literatures that have the appropriate band gap, indicating accurate perovskite composition. We selected the cases where dominant recombination losses are known, and sufficient features are available to assess our model's performance and compare it to the experimental results. We pass the experimental data as an input to our optimal model, and Table S2 (supplemental information) demonstrates a robust agreement between our optimal ML model predictions and the experimental output. Furthermore, majority of experimental findings from existing literatures underscores the recombination losses at interface.[19,24,30] Our optimal model prediction shows a strong agreement with the experimental findings by predicting the dominant recombination losses at interface as depicted in Table S2 (supplemental information). This strengthens our confidence in the practical applicability of our approach for predicting dominant recombination losses in PSCs.

In summary, the dominant losses in the recombination were investigated at the GB, interface, and band-to-band recombination of PSCs using various ML algorithms. Among the various implemented models, the XGBoost classifier accurately predicted the dominant losses with an accuracy score of 85% on the performance evaluation matrix (sensitivity, specificity, MCC, and accuracy score). To ensure the generalizability of the proposed models, a 7-fold cross-validation method was used. For the best hyperparameter optimization, Optuna was applied to the proposed model, and SHAP analysis was implemented to determine the influence of the features. Furthermore, this study provides a viable option for developing efficient and intelligent PSCs with minimal dominant recombination losses.

Limitations of the study

While the study successfully tackles dominant recombination in PSCs through ML models, a noteworthy limitation is the absence of a dedicated computational web server. Such a tool would be instrumental in precisely identifying dominant recombination losses in PSCs, offering an accessible user interface for the material science community.

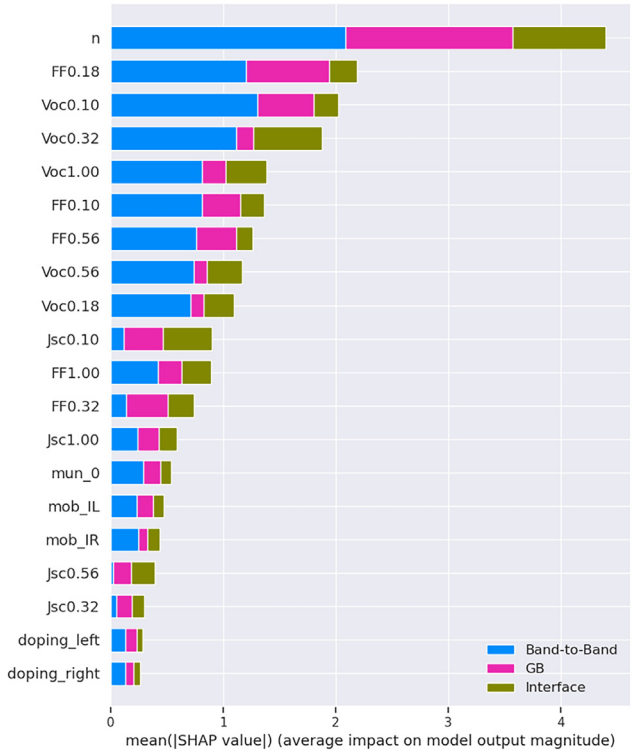


Figure 4. An overview of the SHAP values, showing the 20 most important features for the proposed model for predicting the dominant recombination loss in PSCs

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANTS DETAILS
- METHOD DETAILS
  - Data Set and features
  - Methodology
  - Performance evaluation
  - Hyperparameter optimization
  - Feature importance
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2024.109200.

ACKNOWLEDGMENTS

This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. 2020R1A2C2005612). It was also supported by the "Human Resources Program in Energy Technology" of the Korea Institute of Energy Technology Evaluation and Planning (KETEP) and was granted financial resources by the Ministry of Trade, Industry, and Energy, Republic of Korea. (No. 20204010600470).

AUTHOR CONTRIBUTIONS

B.A.: Conceptualization, data curation, formal analysis, investigation, methodology, software, validation, visualization, writing – original draft, writing – review and editing. H.T.: conceptualization, formal analysis, funding acquisition, investigation, resources, supervision, writing – review and editing. K.T.C.: conceptualization, funding acquisition, investigation, supervision, writing – review and editing.

DECLARATION OF INTERESTS

The authors declare no conflict of interest.

Received: October 3, 2023
Revised: December 12, 2023
Accepted: February 7, 2024
Published: February 12, 2024

REFERENCES

1. Park, N.-G. (2015). Perovskite solar cells: an emerging photovoltaic technology. Mater. Today *18*, 65–72. https://doi.org/10.1016/j.mattod.2014.07.007.
2. Kim, J.Y., Lee, J.-W., Jung, H.S., Shin, H., and Park, N.-G. (2020). High-Efficiency Perovskite Solar Cells. Chem. Rev. *120*, 7867–7918. https://doi.org/10.1021/acs.chemrev.0c00107.
3. Ward, L., Agrawal, A., Choudhary, A., and Wolverton, C. (2016). A general-purpose machine learning framework for predicting properties of inorganic materials. npj Comput. Mater. *2*, 16028. https://doi.org/10.1038/npjcompumats.2016.28.
4. Beam, A.L., and Kohane, I.S. (2018). Big Data and Machine Learning in Health Care. JAMA *319*, 1317–1318. https://doi.org/10.1001/jama.2017.18391.
5. Carleo, G., Cirac, I., Cranmer, K., Daudet, L., Schuld, M., Tishby, N., Vogt-Maranto, L., and Zdeborová, L. (2019). Machine learning and the physical sciences. Rev. Mod. Phys. *91*, 045002. https://doi.org/10.1103/RevModPhys.91.045002.
6. Dixon, M.F., Halperin, I., and Bilokon, P. (2020). Machine Learning in Finance (Springer International Publishing). https://doi.org/10.1007/978-3-030-41068-1.
7. Bhavsar, P., Safro, I., Bouaynaya, N., Polikar, R., and Dera, D. (2017). Machine Learning in Transportation Data Analytics. In Data Analytics for Intelligent Transportation Systems (Elsevier), pp. 283–307. https://doi.org/10.1016/B978-0-12-809715-1.00012-2.
8. Häse, F., Roch, L.M., and Aspuru-Guzik, A. (2019). Next-Generation Experimentation with Self-Driving Laboratories. Trends Chem. *1*, 282–291. https://doi.org/10.1016/j.trechm.2019.02.007.
9. Lampe, C., Kouroudis, I., Harth, M., Martin, S., Gagliardi, A., and Urban, A.S. (2023). Rapid Data-Efficient Optimization of Perovskite Nanocrystal Syntheses through Machine Learning Algorithm Fusion. Adv. Mater. *35*, 2208772.
10. Khan, A., Tayara, H., and Chong, K.T. (2023). Prediction of organic material band gaps using graph attention network. Comput. Mater. Sci. *220*, 112063. https://doi.org/10.1016/j.commatsci.2023.112063.
11. Butler, K.T., Davies, D.W., Cartwright, H., Isayev, O., and Walsh, A. (2018). Machine learning for molecular and materials science. Nature *559*, 547–555. https://doi.org/10.1038/s41586-018-0337-2.
12. Ismail, Z.S., Sawires, E.F., Amer, F.Z., and Abdellatif, S.O. Perovskites

informatics: Studying the impact of thicknesses, doping, and defects on the perovskite solar cell efficiency using a machine learning algorithm. Int. J. Numer. Model. Electron. Networks, Devices Fields. 37 e3164

13. Sanchez-Lengeling, B., Roch, L.M., Perea, J.D., Langner, S., Brabec, C.J., and Aspuru-Guzik, A. (2019). A Bayesian Approach to Predict Solubility Parameters. Adv. Theory Simul. *2*, 1800069. https://doi.org/10.1002/adts.201800069.
14. Wilbraham, L., Sprick, R.S., Jelfs, K.E., and Zwijnenburg, M.A. (2019). Mapping binary copolymer property space with neural networks. Chem. Sci. *10*, 4973–4984. https://doi.org/10.1039/C8SC05710A.
15. Liu, Y., Tan, X., Liang, J., Han, H., Xiang, P., and Yan, W. (2023). Machine learning for perovskite solar cells and component materials: key technologies and prospects. Adv. Funct. Mater. *33*, 2214271.
16. Salah, M.M., Ismail, Z., and Abdellatif, S. (2023). Selecting an appropriate machine-learning model for perovskite solar cell datasets. Mater. Renew. Sustain. Energy *12*, 187–198.
17. Nagasawa, S., Al-Naamani, E., and Saeki, A. (2018). Computer-Aided Screening of Conjugated Polymers for Organic Solar Cell: Classification by Random Forest. J. Phys. Chem. Lett. *9*, 2639–2646. https://doi.org/10.1021/acs.jpclett.8b00635.
18. Lu, Y., Wei, D., Liu, W., Meng, J., Huo, X., Zhang, Y., Liang, Z., Qiao, B., Zhao, S., Song, D., and Xu, Z. (2023). Predicting the device performance of the perovskite solar cells from the experimental parameters through machine learning of existing experimental results. J. Energy Chem. *77*, 200–208.
19. Sherkar, T.S., Momblona, C., Gil-Escrig, L., Ávila, J., Sessolo, M., Bolink, H.J., and Koster, L.J.A. (2017). Recombination in Perovskite Solar Cells: Significance of Grain Boundaries, Interface Traps, and Defect Ions. ACS Energy Lett. *2*, 1214–1222. https://doi.org/10.1021/acsenergylett.7b00236.
20. Wolff, C.M., Caprioglio, P., Stolterfoht, M., and Neher, D. (2019). Nonradiative Recombination in Perovskite Solar Cells: The Role of Interfaces. Adv. Mater. *31*, 1902762. https://doi.org/10.1002/adma.201902762.
21. Correa-Baena, J.-P., Tress, W., Domanski, K., Anaraki, E.H., Turren-Cruz, S.-H., Roose, B., Boix, P.P., Grätzel, M., Saliba, M., Abate, A., and Hagfeldt, A. (2017). Identifying and suppressing interfacial recombination to achieve high open-circuit voltage in perovskite solar cells. Energy Environ. Sci. *10*, 1207–1212. https://doi.org/10.1039/C7EE00421D.
22. Guo, Z., Jena, A.K., Kim, G.M., and Miyasaka, T. (2022). The high open-circuit voltage of perovskite solar cells: a review. Energy Environ. Sci. *15*, 3171–3222. https://doi.org/10.1039/D2EE00663D.
23. Luo, D., Su, R., Zhang, W., Gong, Q., and Zhu, R. (2019). Minimizing non-radiative recombination losses in perovskite solar cells. Nat. Rev. Mater. *5*, 44–60.
24. Le Corre, V.M., Stolterfoht, M., Perdigón Toro, L., Feuerstein, M., Wolff, C., Gil-Escrig, L., Bolink, H.J., Neher, D., and Koster, L.J.A. (2019). Charge transport layers limiting the efficiency of perovskite solar cells: how to optimize conductivity, doping, and thickness. ACS Appl. Energy Mater. *2*, 6280–6287.
25. Calado, P., Burkitt, D., Yao, J., Troughton, J., Watson, T.M., Carnie, M.J., Telford, A.M., O'Regan, B.C., Nelson, J., and Barnes, P.R. (2019). Identifying dominant recombination mechanisms in perovskite solar cells by measuring the transient ideality factor. Phys. Rev. Appl. *11*, 044005.
26. Chen, J., and Park, N. (2019). Causes and solutions of recombination in perovskite solar cells. Adv. Mater. *31*, 1803019.
27. Le Corre, V.M., Sherkar, T.S., Koopmans, M., and Koster, L.J.A. (2021). Identification of the dominant recombination process for perovskite solar cells based on machine learning. Cell Rep. Phys. Sci. *2*, 100346. https://doi.org/10.1016/j.xcrp.2021.100346.
28. Odunmbaku, G.O., Chen, S., Guo, B., Zhou, Y., Ouedraogo, N.A.N., Zheng, Y., Li, J., Li, M., and Sun, K. (2022). Recombination Pathways in Perovskite Solar Cells. Adv. Mater. Interfaces *9*, 1–22. https://doi.org/10.1002/admi.202102137.
29. Castro-Méndez, A., Hidalgo, J., and Correa-Baena, J. (2019). The role of grain boundaries in perovskite solar cells. Adv. Energy Mater. *9*, 1901489.
30. Stolterfoht, M., Wolff, C.M., Márquez, J.A., Zhang, S., Hages, C.J., Rothhardt, D., Albrecht, S., Burn, P.L., Meredith, P., Unold, T., and Neher, D. (2018). Visualization and suppression of interfacial recombination for high-efficiency large-area pin perovskite solar cells. Nat. Energy *3*, 847–854.
31. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., and Dubourg, V. (2011). Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.
32. Chen, T., and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785.
33. Dorogush, A.V., Ershov, V., and Gulin, A. (2018). CatBoost: gradient boosting with categorical features support. Preprint at arXiv. https://doi.org/10.48550/arXiv.1810.11363.
34. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. Adv. Neural Inf. Process. Syst. *30*, 1–8.
35. Lundberg, S.M., and Lee, S.-I. (2017). A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. *30*, 1–9.
36. Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, pp. 2623–2631.
37. Sherkar, T.S., Momblona, C., Gil-Escrig, L., Bolink, H.J., and Koster, L.J.A. (2017). Improving Perovskite Solar Cells: Insights From a Validated Device Model. Adv. Energy Mater. *7*, 1602432. https://doi.org/10.1002/aenm.201602432.
38. Wehrenfennig, C., Eperon, G.E., Johnston, M.B., Snaith, H.J., and Herz, L.M. (2014). High Charge Carrier Mobilities and Lifetimes in Organolead Trihalide Perovskites. Adv. Mater. *26*, 1584–1589. https://doi.org/10.1002/adma.201305172.
39. Koopmans, M., Corre, V., and Koster, L. (2022). SIMsalabim: An open-source drift-diffusion simulator for semiconductor devices. J. Open Source Softw. *7*, 3727.
40. Neukom, M.T., Schiller, A., Züfle, S., Knapp, E., Ávila, J., Pérez-Del-Rey, D., Dreessen, C., Zanoni, K.P.S., Sessolo, M., Bolink, H.J., and Ruhstaller, B. (2019). Consistent device simulation model describing perovskite solar cells in steady-state, transient, and frequency domain. ACS Appl. Mater. Interfaces *11*, 23320–23328.
41. Calado, P., Telford, A.M., Bryant, D., Li, X., Nelson, J., O'Regan, B.C., and Barnes, P.R.F. (2016). Evidence for ion migration in hybrid perovskite solar cells with minimal hysteresis. Nat. Commun. *7*, 13831.
42. Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. Mach. Learn. *63*, 3–42. https://doi.org/10.1007/s10994-006-6226-1.
43. Powers, D.M.W. (2020). Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation, pp. 37–63.
44. Alam, W., Tayara, H., and Chong, K.T. (2020). XG-ac4C: identification of N4-acetylcytidine (ac4C) in mRNA using eXtreme gradient boosting with electron-ion interaction pseudopotentials. Sci. Rep. *10*, 20942. https://doi.org/10.1038/s41598-020-77824-2.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Dataset	Le Corre et al.[27]	https://github.com/kostergroup/Perovskite-Device-Doctor
Software and algorithms		
Scikit learn. 1.2.1	Pedregosa et al.[31]	https://scikit-learn.org/stable/supervised_learning.html#supervised-learning
XGBoost Classifier 1.7.1	Chen et al.[32]	https://xgboost.readthedocs.io/en/stable/
CatBoost Classifier 1.0.1	Dorogush et al.[33]	https://catboost.ai/en/docs/concepts/python-reference_catboostclassifier
LightGBM 3.3.5	Guolin Ke et al.[34]	https://lightgbm.readthedocs.io/en/stable/index.html
Python 3.10	Python Software Foundation	https://www.python.org
SHAP 0.41.0	Lundberg et al.[35]	https://shap.readthedocs.io/en/latest/
Optuna optimization framework 3.1.1	Akiba et al.[36]	https://optuna.org/#key_features

RESOURCE AVAILABILITY

Lead contact

For more details and the necessary requirements concerning additional resources, please direct your inquiries to Dr. Hilal Tayara (hilaltayara@jbnu.ac.kr).

Materials availability

Discovery of unique materials was not part of this study, as it did not involve the use of any distinctive reagents.

Data and code availability

- The data and code are made available on GitHub at https://github.com/BasirAkbar/xgboost_perovskite.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.
- GPU: Nvidia Titan-12 GB RAM.
- Hardware requirements: intel(R) Core (TM) i5-10500 CPU @ 3.10GHz 3.10 GHz.

EXPERIMENTAL MODEL AND STUDY PARTICIPANTS DETAILS

This study leveraged the ML framework for predicting the dominant recombination losses in PSCs. Furthermore, the validity of our optimal ML model was confirmed through experimental data. A robust alignment was observed when comparing the ML model predictions with the experimental outcomes. Consequently, we express confidence in our approach to predict dominant recombination losses in PSCs, as it has proven itself to be a more reliable choice.

METHOD DETAILS

Data Set and features

The dataset was made available on PSCs along with the light intensity-dependent performance and the dominant recombination process (an already known process) for every datum.[27] This has demonstrated the ability to produce PSCs by creating various devices with different compositions and structures.[37] In the dataset details, random parameters, parasitic lumped series, and shunt resistance were selected in a specific range. Therefore, a lumped series resistance was used instead of a distributed series resistance. Moreover, the band gap was fixed at approximately 1.6 eV (MAPbI3).[27] However, the interpretation is incomplete because various parameters that can influence the ideality factor must be combined with other essential parameters, such as Voc,[21] FF, mobilities,[38] Jsc, and doping.[24] In this study, 'doping left' and 'doping right'

represent the doping in both TLs. The assignment of 'left' and 'right' is based upon the arrangement of the HTL and ETL, varying in devices with positive intrinsic negative (pin) or negative intrinsic positive (nip) based device structures. The dataset utilized in this study is applicable to both nip and pin structured devices.[27] Therefore, 'doping left' denotes the doping in the leading transport layer where sunlight penetrates, while 'doping right' refers to the doping in the subsequent transport layer, based on whether it's a nip or pin structured device. The dataset of approximately 2.5 million data points were created by the computational simulations using the open source software SIMsalabim[39] it possesses the capability to conduct both steady-state and time dependent simulations encompassing the various effects such as mobile ions, recombination, trapping and dielectric mismatch31. In this study context, computationally numerical solution of DD equations was employed for various parameters combinations and conditions. DD simulation has been widely applied in PSCs to comprehend their functioning and replicate various devices characterized by perovskite compositions and structures.[24,40] The simulation process involves repetitive computations while adjusting specific parameters that significantly affect the behavior of charges (electrons and holes). These parameters encompass variations in dimensions, thickness, mobilities, doping concentrations, temperature, voltages, or light intensities. These parameters were derived from previously published literature containing the experimental values.[41] The process iterated multiple times using these experimental values alongside different parameter combinations, resulting in a dataset comprising millions of data points. This dataset encapsulates diverse scenarios and expected device conditions, effectively addressing recombination in PSCs. Next, to preprocess the data, the dataset contained (2470491) simulation data. This study predicted three labels: band-to-band, GB, and interface using the given dataset. The dataset was balanced between these labels to ensure that the ML algorithms worked efficiently. Therefore, the dataset was preprocessed to eliminate duplicate and null values from the respective dataset to ensure that the dataset is completely balanced and more accurate. The data were distributed equally among these labels to predict the dominant recombination loss in PSCs. Furthermore, 22 features were selected for training by investigating their correlations with the target variable confirmed through the Pearson correlation coefficient. Furthermore, Feature importance was investigated using a sklearn library and validated using the SHAP analysis method. An equal distribution of target labels was demonstrated in a meticulously selected dataset of 823,497 samples, ensuring a balanced representation. Additionally, the dataset was divided into 75% training and 25% testing datasets. To train the models efficiently, seven-fold cross-validation was used to divide the training data into seven subgroups. Six of them were used for training and one for validation, which helped validate the proposed models and achieve a more generalized model.

Methodology

The overarching goal of this multiclassification task is to precisely identify the input data and accurately predict the final output. The final output encompasses three distinct categories: band-to-band recombination, GB, and interfaces. The primary aim of this study is to devise a proficient model that can discern and classify input data into designated output categories with a remarkable level of accuracy. Consequently, the hyperparameter optimization technique (Optuna) was used to meticulously select the hyperparameters for each implemented model. The ability of the model to precisely differentiate and categorize input samples is crucial for the successful execution of this multiclassification problem. A detailed overview of the ML classification algorithms used in this study is provided in the following sections.

Random Forest

RF is an ensemble-learning algorithm that merges multiple decision trees to make predictions.[31] During the training process, the algorithm randomly selects a subset of the input features and data to train each tree in the forest. This reduces overfitting by training each decision tree on different data subsets. Overall, RF classification is a powerful ML algorithm that can be used to predict categorical output variables using input features.

LightGBM

LightGBM is an open-source gradient boosting framework that uses a tree-based learning algorithm. The architecture constructs the decision tree level-by-level, imitating the approach commonly employed by other gradient boosting algorithms.[34] To obtain a robust and highly optimized LGBM classifier with carefully selected parameters, such as a boosting type of 'gbdt,' a 'multiclass' objective and 'multi-logloss' metric are critical aspects of this research. It allows effective classification and prediction of complex data, minimizes overfitting, and maximizes accuracy. Furthermore, the incorporation of seven-fold cross-validation enhances the robustness and reliability of the model by evaluating its performance across a diverse range of datasets. This approach provides a comprehensive assessment of the model's effectiveness. Additionally, the inclusion of the feature fraction (0.5828) and bagging fraction (0.8719) collectively contributed to the models' capacity to capture intricate details within the data. These measures effectively mitigate overfitting and enhance the ability of the model to generalize unseen data, which boosts its overall performance and predictive capability. More detail about LightGBM can be seen in ML models description (supplemental information).

Extra tree classifier

The extra-tree classifier combines multiple results to make predictions and performs the final classification to construct multiple decision trees.[42] To ensure the models' ability to efficiently generalize new and unseen data using seven-fold cross-validation, the input features and their corresponding labels were used during the model training phase with an extra-tree classifier.

CatBoost

The CatBoost classifier was trained on the dataset using seven-fold cross-validation. It uses a series of decision trees and special hyperparameters to optimize its performance.[33] The hyperparameters included the number of iterations (2128), learning rate (0.082), verbose level (129), tree depth (16), L2 regularization parameter (5), random seed (76), number of iterations required to wait for the optimal solution (200), and metric period (129). These hyperparameters were carefully selected to ensure the models' ability to generalize new data and make accurate predictions. Overall, the CatBoost classifier is a powerful and flexible algorithm that delivers excellent accuracy and provides accurate predictions. Therefore, the CatBoost classifier is a powerful and flexible algorithm. By carefully selecting hyperparameters and using techniques such as cross-validation, excellent accuracy and accurate predictions can be achieved.

XGBoost

The XGBoost algorithm sequentially constructs a series of decision trees and combines their predictions to make a final classification.[32] The XGBoost algorithm is a valuable tool for accurate and reliable classification when dealing with multiclassification problems. XGBoost outperforms the above-mentioned models in terms of accuracy and predictive power. The XGBoost model was trained using a robust approach known as seven-fold cross-validation, which ensures the generalizability of the model to the unseen data. The XGBoost model is further enhanced using complex hyperparameters, such as the number of estimators, maximum depth, minimum child weight, learning rate, regularization lambda, subsampling rate (subsample), number of parallel threads, and maximum number of bins. These hyperparameters were carefully selected and fine-tuned to optimize the performance of the model and enable it to effectively capture complex patterns in the data. Tuning the XGBoost hyperparameters can be challenging because of their interdependencies. To address this challenge, advanced techniques, such as Optuna, were used to automatically search for the best combination of hyperparameters. Optuna employs optimization algorithms to efficiently explore the hyperparameter space and determine the optimal configuration for maximizing model performance. SHAP was used to predict the XGBoost model and identify the key features that contribute to model decisions. It provides valuable insights into the underlying patterns and relationships in the data, helping to guide further analysis and decision-making. Overall, XGBoost can be considered an optimized model to yield maximum accuracy compared with all models implemented in this study. Finally, it provides the best prediction for the classification task compared with the other models. More detail about XGBoost can be seen in ML models description (supplemental information). The implementation details of the ML models utilized in this study can be found in the supplemental information.

Performance evaluation

The performance of the classification model was investigated using the MCC, sensitivity, specificity, and accuracy scores to determine the validity of the model[43,44] and can be measured by the equation given below:

$$MCC = \frac{(tp * tn - fp * fn)}{\sqrt{(tp+fp)(tp+fn)(tn+fp)(tn+fn)}}$$

$$sensitivity = \frac{tp}{tp+fn}$$

$$specificity = \frac{tn}{tn+fp}$$

$$accuracy = \frac{tp+tn}{tp+tn+fp+fn}$$

where tp, tn, fn, and fp denote true positive, true negative, false negative, and false positive, respectively. The detail of the evaluation metrics can be seen in supplemental information.

Hyperparameter optimization

The optimal hyperparameters were selected using the Optuna hyperparameter optimization technique.[36] Optuna is a specialized software framework for automatic hyperparameter optimization in ML. It leverages the Bayesian optimization method, which was specifically designed for fine-tuning the hyperparameters in ML models. Optuna was extensively used to exhaustively explore all potential hyperparameter combinations to achieve exceptional performance on our unique dataset. The objective was to identify the ideal hyperparameter configurations to optimize the performance of our model. Thus, to effectively use this hyperparameter optimization method, we applied this approach to the XGBoost classifier, and all models implemented during this study. This allowed us to achieve the best hyperparameter combination, which directly influenced the performance of the proposed model on the target variable. Moreover, insufficient hyperparameter tuning can result in suboptimal performance in ML and deep-learning models. Hyperparameter optimization is critical for determining the output and overall effectiveness of an ML model. Table S3 (supplemental information) shows the optimal hyperparameter settings derived from Optuna optimization. It provides details of the XGBoost model, search space, and optimal combination achieved. The best hyperparameter combinations

is used to train the proposed model, which tends to achieve an accuracy score of 85%. Tables S4–S6 (supplemental information) list the respective models, search space, and optimal combinations achieved using Optuna.

Feature importance

This section discusses how each feature affects the prediction of the target variables. SHAP was implemented to assess the importance and contribution of the features.[35] A SHAP analysis methodology is similar to that of a parametric analysis in that variables are altered while other variables are kept constant to observe the impact of varying variables on the target variable. The XGBoost model was selected for the SHAP analysis because it outperformed the other models in terms of accuracy. Initially, we evaluated the significance of the input variables to understand their influence on the prediction of the target variable. Figure 4 shows the importance of each input variable. The SHAP analysis provided compelling evidence of the robustness of the XGBoost classifier's performance in accurately predicting the dominant recombination in PSC. Notably, the ideality factor (N), FF (0.18), V_{OC} (0.10), and V_{OC} 0.32 emerged as highly influential features for determining the target variable. These features demonstrate that they are significantly more important in predicting all three classes than the other features, as shown in Figure 4. Comprehensive SHAP analysis provides invaluable insights into the relative significance and contribution of each feature in predicting the target variable, thereby enhancing our understanding of the underlying relationships within the dataset. Additionally, Figures S9–S11 (supplemental information) outline separate SHAP analysis for individual classes, highlighting the importance of features for each class.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical analysis were employed to assess and identify the most optimal ML model in our study. Various performance evaluation metrics were utilized, and the models were meticulously compared. Additionally, Receiver Operating Characteristic (ROC) curves were generated for each model to scrutinize their discriminatory capabilities between positive and negative classes. Furthermore, a classification report was generated for each ML model to provide a comprehensive assessment of their overall performance in this study.