Computer Modeling in Engineering & Sciences

Tech Science Press

Doi:10.32604/cmes.2025.0xxxx

ARTICLE

Optimizing performance prediction of perovskite photovoltaic materials by statistical methods-intelligent calculation model

Guo-Feng Fan[1,3], Jia-Jing Qian[1], Li-Ling Peng[1], Xin-Hang Jia[1], Ling-Han Zuo[1], Jia-Can Yan[1], Jiang-Yan Chen[1], Anantkumar J. Umbarkar[4], and Wei-Chiang Hong[2,*]

[1]School of Mathematics & Statistics, Ping Ding Shan University, Ping Ding Shan 467000, Henan, China
[2]College of Shipbuilding Engineering, Harbin Engineering University, Harbin, 150001, China
[3]Yaoshan Lab, Ping Ding Shan 467000, Henan, China
[4]Department of Information Technology, Walchand College of Engineering, Sangli, Maharashtra, 416415, India
[*]Corresponding Author: Wei-Chiang Hong. Email: samuelsonhong@gmail.com

Received: 12 September 2025; Accepted: Day Month Year; Published: Day Month Year

ABSTRACT: Accurate prediction of perovskite photovoltaic materials' optoelectronic properties is crucial for developing efficient and stable materials, advancing solar technology. To address poor interpretability, high computational complexity, and inaccurate predictions in relevant machine learning models, this paper proposes a novel methodology. The technical route of this paper mainly centers on the random forest-least absolute shrinkage and selection operator regression enhanced knowledge distillation-bidirectional gated recurrent unit with attention technology (namely RFLR-EKD-BIGRUA), which is applied in perovskite photovoltaic materials. Primarily, it combines random forest and Lasso regression to quantitatively assess feature importance, selecting variables with significant impacts on photoelectric conversion efficiency. Subsequently, statistical techniques analyze the weight distribution of variables influencing power conversion efficiency (PCE) to extract key features. In the model optimization phase, knowledge distillation transfers features from complex teacher models to student models, enhancing prediction accuracy. Additionally, BiGRU-Attention is introduced to further optimize predictive performance while substantially reducing computational costs. Results show that integrating statistical techniques into intelligent optimization models quantifies photovoltaic system uncertainties, minimizes prediction errors, enables efficient screening of perovskite materials meeting energy storage standards, and offers precise material selection guidance.

KEYWORDS: Perovskite photovoltaic cells; Random forest; Lasso regression; Knowledge distillation; BiGRU-Attention

1 Introduction

The intensification of energy shortage problems has accelerated the global push for clean energy technological innovation, spurring the rapid development of photovoltaic technology for the efficient conversion of renewable energy. Among these innovations, perovskite photovoltaic materials [1] have become a focal point in

the field of photovoltaics due to their high efficiency and low cost. Nevertheless, the present paucity of interpretability in machine learning (ML) models [2] has resulted in suboptimal efficiency in the selection of perovskite photovoltaic battery materials. Consequently, the screening of significant feature variables to enhance the performance of predictive models, reduce prediction errors, alleviate the experimental burden, and improve selection efficiency represents a pivotal direction and opportunity for China's photovoltaic industry [3].

1.1 Literature Review

It is anticipated that the market for perovskite solar cells will expand quickly due to the continuous global research and development of renewable energy technologies. The use of this clean energy technology is expected to become more and more important in lowering the carbon footprint of the world [4]. Chalcogenide photovoltaic (PV) cells, on the other hand, have become a major area of study because of their proven high photoelectric conversion efficiency. Historically, the selection of high-performance perovskite photovoltaic materials [5-6] has relied on lengthy and often inefficient experimental processes that lacked systematic methods for predicting and optimizing material properties. However, with the substantial accumulation of both experimental and computational data [7], alongside the gradual maturation of ML techniques, this led to the notable integration of ML approaches into the development of perovskite photovoltaic materials. These studies use ML to conduct comprehensive analyses regarding how characteristic variables influence PCE, as well as to predict PCE. This facilitates targeted material design and optimization [8-9]. In this process, statistical analysis plays a crucial role. It not only provides a data foundation for ML models but also helps researchers extract valuable information from massive datasets, thereby enabling more accurate predictions of material properties. In the field of predicting perovskite photovoltaic material performance, researchers have utilized diverse ML models, including random forest (RF) [10-12], support vector machine (SVM) [13-14], and LASSO [15-16]. Tao et al. [17] compiled and analyzed applications of ML in large-scale screening and rational design of perovskite photovoltaic materials, offering insights into the prospects for ML within this field. Kusuma et al. [18] The authors employed ML techniques to analyze the structure of high-efficiency perovskite solar cells (PSCs), revealing new, efficient configurations that improve device performance. Meanwhile, Mishra et al. [19] selected perovskite materials with suitable band gaps and then simulated and analyzed the performance of corresponding devices under indoor lighting conditions. Yeom et al. [1] engineered multi-quantum well perovskites with long organic spacers and oleylamine-treated interfaces, enhancing radiative efficiency and stability. In predicting device performance parameters, Lu et al. [20] employed ML-guided inverse design for PSCs, reducing experimental redundancy while accelerating efficient material screening via pre-experimental statistical analysis. This approach provides a vital foundation for the ML-based prediction of perovskite photovoltaic material performance PCE before experimentation. Alfares et al. [21] applied Bayesian-optimized ML to multi-feature perovskite data (ionic/bulk properties), enhancing lattice constant prediction accuracy. Furthermore, a number of studies have examined the statistical applications of ML technology [22-24]. These investigations demonstrate that statistical analysis can aid researchers in extracting valuable insights from limited data, while also establishing a scientific basis for model optimization and validation. In examining the hole transport layer (HTL) in carbon-based perovskite solar cells (C-PSCs), Valsalakumar et al. [25] verified that the application of

ML has been instrumental in streamlining the optimization process, reducing dependence on traditional trial-and-error methodologies, and facilitating a deeper understanding of the interdependencies among critical device parameters. These investigations further highlight the importance of statistical methods in optimizing material performance. Nevertheless, there are still several key challenges, such as improving the efficiency and accuracy of statistical analyses within optimization processes and utilizing statistical methods more effectively to tackle multi-parameter optimization issues.

As stated in the aforementioned foreign papers, ML algorithms play a crucial role in the research of perovskite solar cell materials and battery structures. To improve the efficiency and stability of perovskite solar cells, researchers have focused on exploring a series of parameters, including the thickness of the absorption layer and the materials used for the HTL and electron transport layer (ETL).

Many researchers have combined different intelligent models with ML algorithms to enhance predictive accuracy and efficiency. For example, Tian et al. [26] developed a transfer learning model for perovskite thickness prediction, overcoming data scarcity by leveraging diverse bandgap material datasets to enhance accuracy. Mannodi-Kanakkithodi and Chan [27] employed DFT to study halide perovskite impurities; integrating these results with ML and statistics enables rapid identification of optically active defects. These impurities can be introduced deliberately to modulate perovskite conductivity and optimize photovoltaic absorption performance. Additionally, Bak et al. [28] developed a k-fold cross-validated deep learning approach to rapidly optimize tin-based perovskite solar cell (SnPSC) structures. This approach leverages limited experimental data on SnPSC while maximizing prediction accuracy through rigorous statistical optimization techniques. While existing studies demonstrate robust statistical validation, advancing accuracy and reliability remains crucial, particularly for small datasets and multi-parameter optimization. Further research is needed to refine these statistical approaches.

The Ayad et al. [29] team has systematically studied the influence mechanisms of absorber layer thickness, defect density, and temperature on battery performance, providing a preliminary research foundation for the experimental construction of the proposed structure. The selection of this structure is based on three major advantages: higher energy conversion efficiency, simple deposition process, and market-accessible low-cost material system. In contrast, our team has deeply explored the action laws of 14 key material parameters, such as high open-circuit voltage (Voc), short-circuit current density (Jsc), fill factor (FF), and the thickness of the compatible HTL on the PCE. The core advantages of the selected structure lie in: Minimizing efficiency loss through narrow-bandgap material screening, optimizing the effective area of the battery, and controlling the thickness of the back ETL; Integrating the Lasso feature selection and Bayesian optimization strategies, and relying on large-scale model integration to construct a closed-loop iterative development system, providing a dynamic optimization path for material reverse engineering. This study not only quantifies the correlation between material parameters and PCE but also deeply couples theoretical research with application requirements through engineering design, establishing a multi-dimensional evaluation framework for the efficient screening of perovskite photovoltaic materials.

Taken together, despite the substantial progress that has been made in researching perovskite photovoltaic materials to date, there are still several urgent challenges that require immediate attention. There is

a need to enhance the integration of experimental and computational data to improve the accuracy of predictive models, to develop ML algorithms specifically designed for small-sample datasets to mitigate limitations arising from restricted experimental data availability, and to further optimize ML models to improve both prediction efficiency and accuracy.

1.2 Research Motivation and Innovation

Currently, challenges remain in the performance prediction and optimization of perovskite photovoltaic cells, which hinder the accuracy and efficiency of PCE prediction for perovskite photovoltaic materials as well as the selection and optimization of these materials. Primary among the challenges is the influence of multiple characteristic factors on photovoltaic conversion efficiency, which significantly compromises model prediction accuracy. Concurrently, the inherent uncertainty of PV systems contributes to prediction inaccuracies and heightened error susceptibility. Furthermore, prevailing machine learning models face dual limitations of inadequate interpretability and computational complexity, ultimately diminishing experimental efficiency while increasing operational costs.

1. To address the challenges posed by the high dimensionality of characteristic variables and the unclear underlying mechanisms in photovoltaic material performance prediction, we propose a collaborative "data-algorithm" dual-driven analytical framework. Random forest and Lasso regression models were integrated with statistical validation methods, including 10-fold cross-validation, to develop a robust feature importance ranking-based variable selection mechanism. This approach enables the precise identification of key factors governing PCE.
2. To overcome the persistent accuracy limitations of lightweight models, we propose a knowledge distillation framework. By developing a teacher-student knowledge transfer framework incorporating a BiGRU-Attention temporal feature extraction module, this method significantly improves prediction accuracy while maintaining low parameter complexity, thereby effectively addressing the fundamental trade-off between model compression and performance preservation.
3. To tackle the critical industrial challenges in new material discovery, particularly prolonged development cycles and excessive trial-and-error costs, we developed a tripartite intelligent research and development system integrating statistical analysis, machine learning, and model distillation. This integrated framework utilizes variable importance analysis to optimize experimental design while leveraging high-performance predictive models for accelerated material screening. Practical implementations demonstrate substantial reductions in both development cycles and empirical optimization costs, thereby offering a robust technical foundation for the commercialization of perovskite photovoltaic materials.

The rest of this paper is as follows. The second part presents the methodology of the model. The third part discusses in detail the complex variables of perovskite PV materials and the application of the coupled intelligent model in predicting photoelectric performance. The method is also verified through experiments. The fourth part summarizes the research content of this paper.

2 Methodology

2.1 Random forest

The random forest algorithm efficiently reduces dimensionality by evaluating feature importance, enabling rapid screening and elimination of redundant features to enhance both model performance and efficiency. In this study, the prediction accuracy and efficiency of the model were improved by identifying and removing feature variables that have minimal impact on model prediction performance. The random forest algorithm consists of steps such as sampling, tree building, repetition, and aggregation, specifically as follows:

Step 1: Construct decision trees. Multiple decision trees are trained based on sub-datasets obtained through random sampling. Each tree is generated from an independent sample subset drawn with replacement from the original dataset. For each sample set, the mean squared error (MSE) of its out-of-bag (OOB) data needs to be calculated using Eq. (1).

$$MSE_{OOB} = \frac{1}{n_{OOB}} \sum_{i \in OOB} (y_i - \hat{y}_i)^2 \quad (1)$$

where n_{OOB} represents the number of OOB data points, y_i is the true value of the ith data point, and $\hat{y}_i$ is the predicted value of the *i*-th data point by the model.

Step 2: Random permutation of feature variables. For each tree in the random forest, select one feature and randomly permute the corresponding feature of its OOB data samples to generate a new test dataset. Then, use the random forest model to make predictions and calculate the new MSE of the OOB data.

Step 3: Calculate feature importance. The importance of each feature variable is determined by comparing the changes in MSE of the OOB data before and after permutation. For each feature X_j, its importance $I(X_j)$ is given by Eq. (2),

$$I(X_j) = \frac{1}{K} \sum_{k=1}^{K} (MSE_{OOB,k} - MSE^*_{OOB,k}) \quad (2)$$

where K is the total number of trees, and $MSE^*_{OOB,k}$ is the MSE of the OOB data for the k-th tree after permuting the feature X_j.

Step 4: Feature comparison and ranking. The importance scores of each feature are normalized for comparison and ranking. The normalized feature importance $\tilde{I}(X_j)$ is given by Eq. (3),

$$\tilde{I}(X_j) = \frac{I(X_j) - min(I)}{max(I) - min(I)} \quad (3)$$

where $max(I)$ and $min(I)$ are the maximum and minimum importance scores of all features, respectively.

The random forest algorithm efficiently reduces dimensionality by evaluating feature importance, enabling rapid screening and elimination of redundant features to enhance both model performance and efficiency. However, the importance of some feature variables is not obvious, which may generate significant noise in regression and cause higher errors in model accuracy. Therefore, LASSO regression should be used to re-screen important features through L1 regularization, reduce noise, provide more concise input features, lower computational complexity, and thereby improve the efficiency of subsequent models. Its working principle is shown in the following Fig. 1.

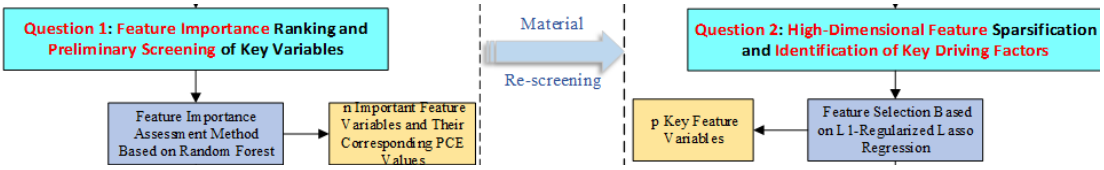

Figure 1: The connection between random forests and Lasso regression

2.2 Least absolute shrinkage and selection operator regression

LASSO regression (Least Absolute Shrinkage and Selection Operator Regression), as a classic algorithm for high-dimensional feature space compression estimation, achieves sparse modeling of the feature space while estimating parameters by introducing an L1-norm penalty term. In this study, based on the glmnet framework in the R language ecosystem, a performance prediction model for perovskite photovoltaic materials was constructed using the Lasso regularization mode. Under the criterion of minimizing the MSE through ten-fold cross-validation, core variables with significant regulatory effects on PCE were selected from 14 candidate features, including carrier transport properties, interfacial engineering parameters, and band structure indicators. Its mathematical essence is given by Eq. (4),

$$\min_{\beta}\left\{\frac{1}{2N}\sum_{i=1}^{N}\left(y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|\right\} \tag{4}$$

where β represents the regression coefficients, N is the number of samples, y_i is the observed value of the i-th sample, β_0 is the intercept term, x_{ij} is the j-th feature value of the i-th sample, λ is the regularization parameter, which controls the model complexity and prevents overfitting, and p is the number of features.

2.3 K-fold cross-validation

K-fold cross-validation splits the dataset multiple times into training and testing sets according to a predetermined proportion, ensuring that all data can alternately participate in model training and validation. This fully exploits the value of limited-sample datasets and significantly enhances the model's generalization performance. As a special case of k-fold cross-validation, the principle of ten-fold cross-validation can be intuitively demonstrated in Fig. 2.

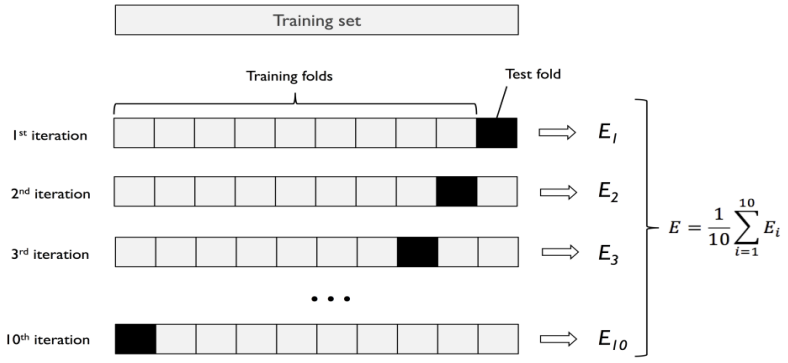

Figure 2: The principle of ten-fold cross-validation

2.4 Knowledge distillation optimization

Knowledge distillation is a process that involves the transfer of knowledge from a complex model (the teacher model) to a simpler model (the student model). The objective of this process is to maintain or enhance performance while reducing model complexity. In this study, the Knowledge distillation optimization algorithm was employed to intelligently enhance the performance of the Random forest model. The enabling of the student model to learn the features of the teacher model has been demonstrated to result in a more accurate capture of the intrinsic structure of the data. This, in turn, has been shown to determine the optimal combination of model parameters and improve the algorithm's predictive capabilities on both the training and test sets, as well as the model's generalization ability.

2.5 The BiGRU-Attention

The BiGRU-Attention model is a deep learning model that integrates Bidirectional Gated Recurrent Units (BiGRU) with the Attention mechanism, rendering it particularly suitable for processing sequential data. BiGRU, a variant of a recurrent neural network (RNN), possesses a bidirectional structure that facilitates the capture of long-term dependencies in temporal data. Concurrently, the Attention mechanism facilitates the model's concentration on the salient components of the input data, thereby enhancing the precision of predictions. To compensate for the distillation loss of logical units in the network output from the teacher to the student end, this study proposes the research idea of the BiGRU-Attention model, with specific steps as follows:

Step 1: The process of data preprocessing is an essential step in the data science workflow. The raw data are then transformed into a format that can be processed by the model. The dataset is divided into training and testing sets at a ratio of 80% and 20%, respectively.

Step 2: The definition of the distillation loss function is given here.

$$L_{KD}(p(u,T),p(z,T)) = \sum_{i=0}^{k} -p_i\,(u_i,T)\,log(\,p_i(z_i,T)) \tag{5}$$

where L_{KD} represents the knowledge distillation loss function, $p(u,T)$ and $p(z,T)$ denote the probability distributions of the teacher and student models at temperature parameter T, respectively, k is the total number of classes, and $p_i(u_i,T)$ and $p_i(z_i,T)$ are the probabilities of the teacher and student models for the ith class.

Step 3: Training the teacher model on the target dataset. The BiGRU-Attention architecture is designed to optimize the initial parameters of the student model. Only the parameters of the student model are updated; the parameters of the teacher model are fixed. Both the task-specific loss and the distillation loss are computed simultaneously during the forward pass; the student model's parameters are optimised during the backward pass.

2.6 Methodological framework

The above content informs the research idea of the RF-Lasso-KD-BiGRU- Attention model proposed by this study.

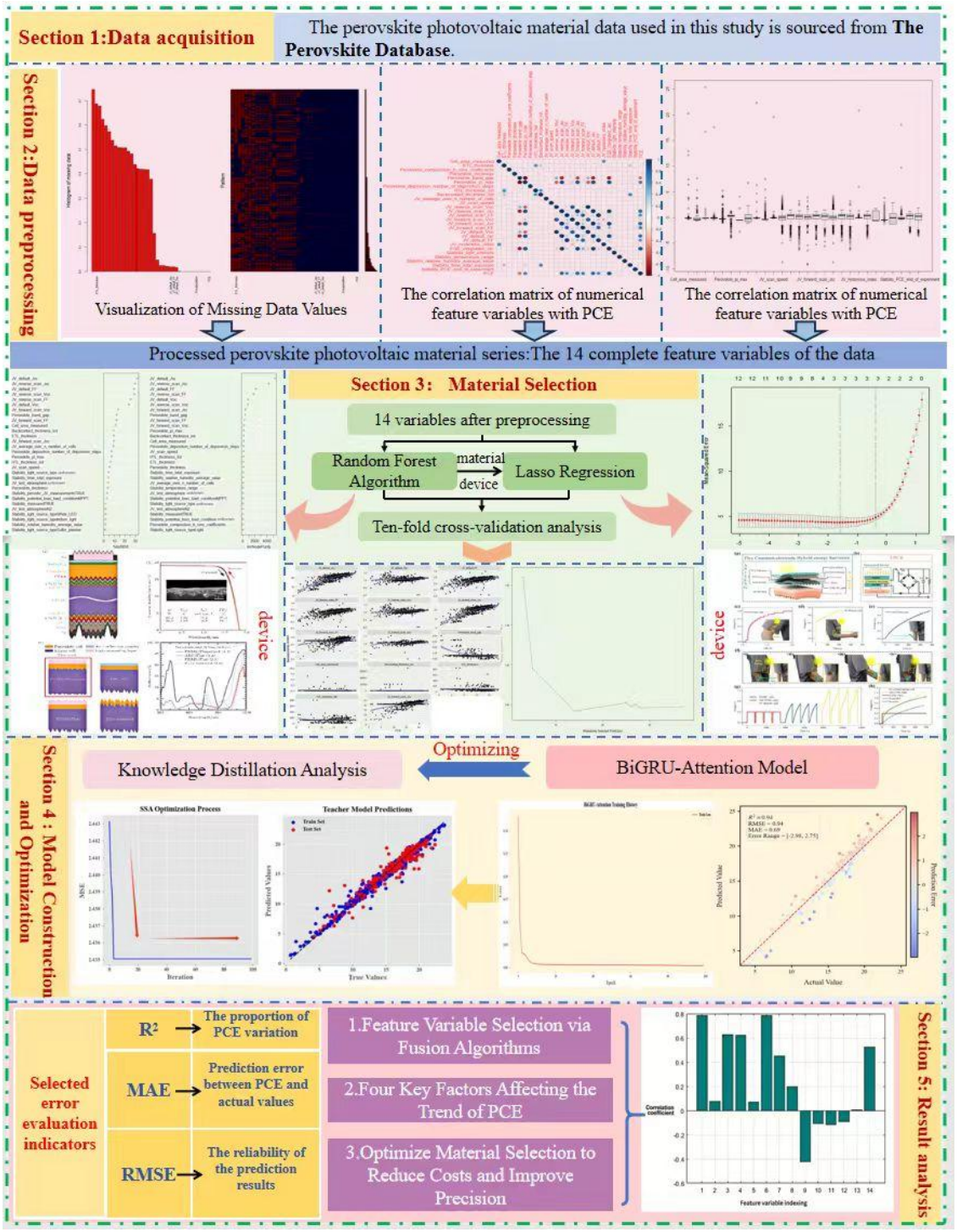

Figure 3: Technical flowchart of the RFLR-EKD-BIGRUA in perovskite photovoltaic materials

Step 1: Data Preprocessing. The historical dataset of perovskite photovoltaic materials is meticulously curated from the Perovskite Database. Through rigorous preprocessing protocols encompassing data cleansing, missing value treatment, and standardization procedures, a high-quality dataset is established to facilitate reliable subsequent investigations.

Step 2: Feature Variable Screening and Focusing. Using the random forest algorithm, n high-importance features are selected from the initial m variables. Subsequently, Lasso regression further refines these features, reducing them to *p* key variables. This two-stage feature selection process effectively transitions from a broad feature space to a precise and interpretable subset.

Step 3: Preliminary Model Construction and Optimization. Based on the selected key features, a random forest regression model is initially constructed to predict the performance of perovskite photovoltaic materials. The model is then optimized using ten-fold cross-validation to enhance its generalization capability.

Step 4: Model Complexity Optimization. The knowledge distillation technique is implemented to transfer learning from the ten-fold cross-validated optimal model to a more compact architecture, achieving parameter efficiency without significant performance degradation.

Step 5: Deep Model Optimization and Perfection. Taking the BiGRU-Attention model as the teacher model, the student model in knowledge distillation is deeply optimized by minimizing the total loss to obtain a high-precision and high-performance prediction model.

The modelling process is illustrated in Fig. 3.

3 Case analysis

3.1 Data sources and explanation

The Perovskite photovoltaic material data employed in this study were sourced from The Perovskite Database [30], a comprehensive repository containing detailed characterization parameters for perovskite photovoltaic materials. The database encompasses various characteristics of perovskite photovoltaic materials, including: fundamental cell parameters, substrate materials, properties of both compact and mesoporous n-type electron transport layers (ETLs), perovskite characteristics, *p*-type HTL properties, back contact electrodes, encapsulation materials, as well as device characteristics during current density-voltage (J-V) measurements, external quantum efficiency (EQE) measurements, and stability tests. This rich dataset enables researchers to employ ML techniques to identify subtle patterns and correlations that emerge only in large-scale analyses. By facilitating data-driven discoveries, the database not only advances current perovskite research but also establishes a robust foundation for future technological developments in this field.

3.2 Data processing

Before commencing the research, data preprocessing must be meticulously executed, encompassing missing value imputation, outlier detection and remediation, and standardized data normalization procedures.

Missing value treatment: Due to excessive missing values in some battery sample data within the dataset, samples with substantial missing values were removed, ensuring that each sample retained the corresponding photovoltaic conversion efficiency (PCE). This process yielded a dataset comprising 981 perov-

skite photovoltaic cell samples. For visualization of missing values, feature variables with correlation coefficients below 0.1 were assumed to be irrelevant and excluded from further analysis. Missing values in the remaining relevant features were imputed using the mean values of their respective columns.

Outlier handling: To prevent abnormal data from affecting model training performance and increasing prediction errors, this study employs the boxplot method to process outliers in the dataset. During the treatment, outliers are handled in the same manner as missing values, with replacement by the mean values of their corresponding feature variables. This approach ensures data integrity and the accuracy of subsequent model construction.

Data standardization: Before modeling with machine learning algorithms, to reduce discrepancies between feature variables of different dimensions, ensure equal contribution of each feature variable in the model, and improve model convergence speed and prediction accuracy, data standardization was adopted for preprocessing input information, as shown in Eq. (6).

$$Z = \frac{(x - \mu)}{\sigma} \tag{6}$$

where x represents the original data of the feature variable, μ denotes the mean value, and σ stands for the standard deviation.

Standardized data facilitates comparison and processing because all features now share the same scale. Each data point reflects the magnitude of the data relative to the mean: values greater than 0 indicate higher than the mean, values less than 0 signify lower than the mean, and a value of 0 means the data point equals the mean.

3.3 The original random forest

In terms of feature variable ranking and preliminary screening of key variables, this paper selected 36 numerical or categorical features, including the effective area of Perovskite photovoltaic cells, cell structure, ETL, Perovskite absorption layer, HTL, back contact electrode, polycrystalline characteristics, encapsulation conditions, and various characteristics measured under three different conditions of J-V, EQE, and stability, as well as the corresponding PCE values. The preprocessed data of 14 feature variables were used as inputs, and the corresponding PCE values were used as outputs to construct a random forest model for measuring feature variable importance and performing feature importance analysis. The overall explanatory rate of all feature variables used for regression on the variance of the target variable PCE was 89.7%. There is a close correlation between the feature variables of Perovskite solar cell materials and PCE values. Fig. 4 shows the results of the feature importance analysis by the random forest algorithm.

The importance of some feature variables is not obvious, which may generate significant noise in regression and introduce high errors to the model accuracy. Therefore, based on the previously constructed random forest regression model, we evaluate the importance of feature variables and rank them accordingly. A selection is made from the features with higher rankings, while those with lower contributions are eliminated. In Fig. 4, the "%IncMSE" denotes the average extent of model performance degradation when a feature is excluded from each tree in the random forest. A higher value of this metric indicates a greater

significance of the feature. The “IncNodePurity” is the total amount of purity increased by the feature in all node splits of the tree. It is typically measured by Gini impurity. A higher value of this indicator indicates that the feature contributes more to reducing the uncertainty of the dataset.

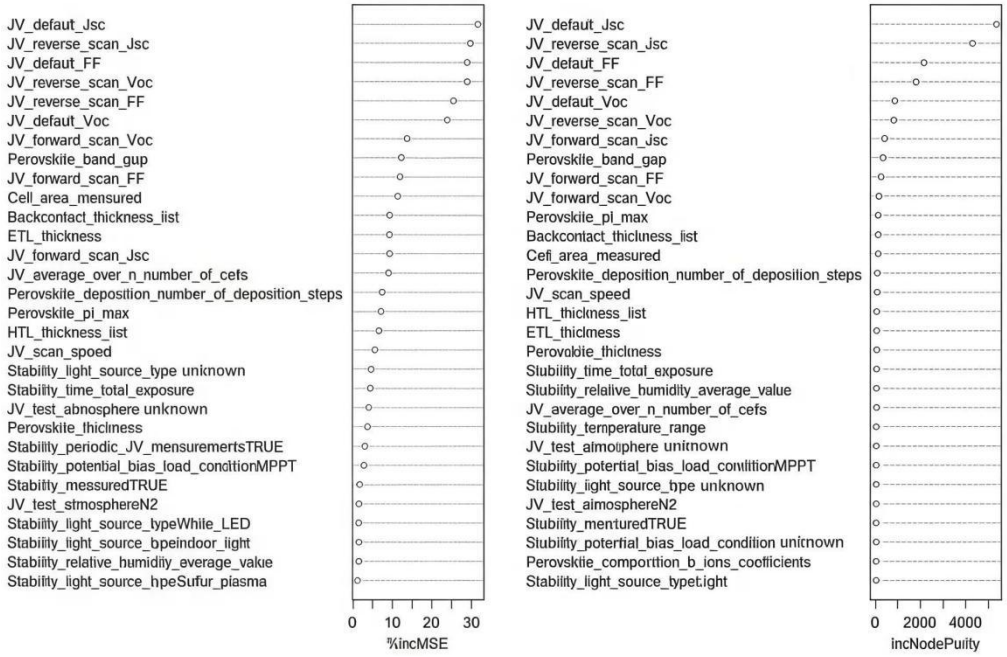

Figure 4: The feature importance analysis in RF

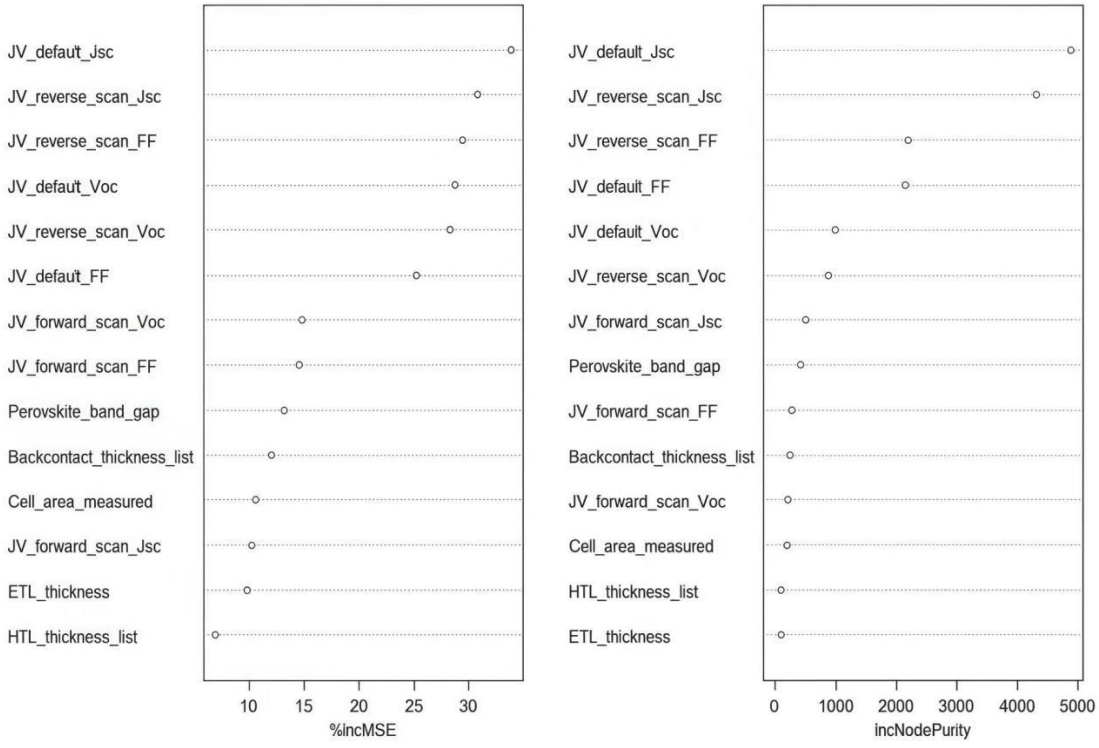

Figure 5: Random forest model with input of the first 14 highly important feature variables

Based on the importance ranking under the "%IncMSE" metric, the top 14 variables were selected to build the final random forest model (Fig. 5). These variables include: Voc, Jsc, FF, bandgap of the Perovskite, effective area of the cell under illumination, HTL, ETL, and the sum of the thicknesses of each layer in the back contact stack, establish the final random forest model. As shown in Fig. 5, the overall explanation rate of the variance related to the PCE of Perovskite photovoltaic materials by 14 predictor variables has reached 78.64%.

3.4 Lasso regression analysis

This study employed Lasso regularization to construct a performance prediction model for Perovskite photovoltaic materials. Through a ten-fold cross-validation approach utilizing MSE minimization criteria, four key variables exhibiting significant regulatory effects on PCE were identified from a pool of 14 candidate features. These features encompass carrier transport properties, interface engineering parameters, and band structure indicators: JV_default_Jsc, JV_default_FF, JV_reverse_scan_Jsc, and JV_forward_scan_FF. The results of cross-validation for the Lasso regression model applied to the prediction of Perovskite photovoltaic materials are presented in Fig. 6 and Fig. 7. The model reached optimal generalization capability (cross-validated MSE=1.38) at λ=0.029. These results demonstrate that the PCE of Perovskite devices is synergistically regulated by the complete carrier generation-transport-collection process parameters.

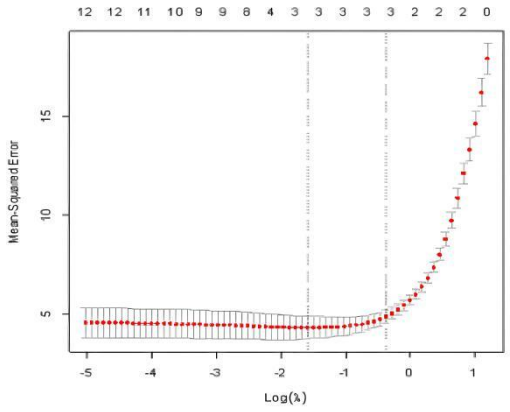

Fig.ure 6: Cross-validation results for the prediction of perovskite photovoltaic materials using the Lasso regression model

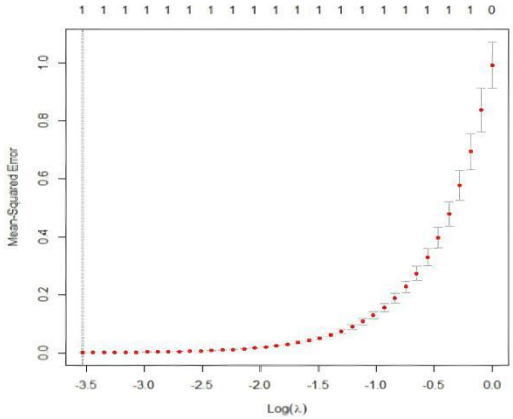

Figure 7: The model performs under different λ values.

3.5 The optimized random forest

In the analysis and optimization of Perovskite photovoltaic material performance, a predictive evaluation model was developed to establish a relationship between 14 key feature variables and PCE. This ultimately led to the construction of a random forest regression model for predicting features and performance. The performance evaluation of the established random forest prediction model yielded the following

results: R^2=0.896, MAE=0.832, and RMSE=1.38. These metrics demonstrate that the final random forest prediction model achieves excellent data fitting accuracy with minimal actual prediction errors.

3.6 K-fold cross-validation analysis

This study employed 10-fold cross-validation to optimize the model's generalization capability with limited data by selecting the most informative feature variables. As shown in Fig. 8, the cross-validation results definitively indicate that the model achieves minimum RMSE and MAE when incorporating the top 14 most important features.

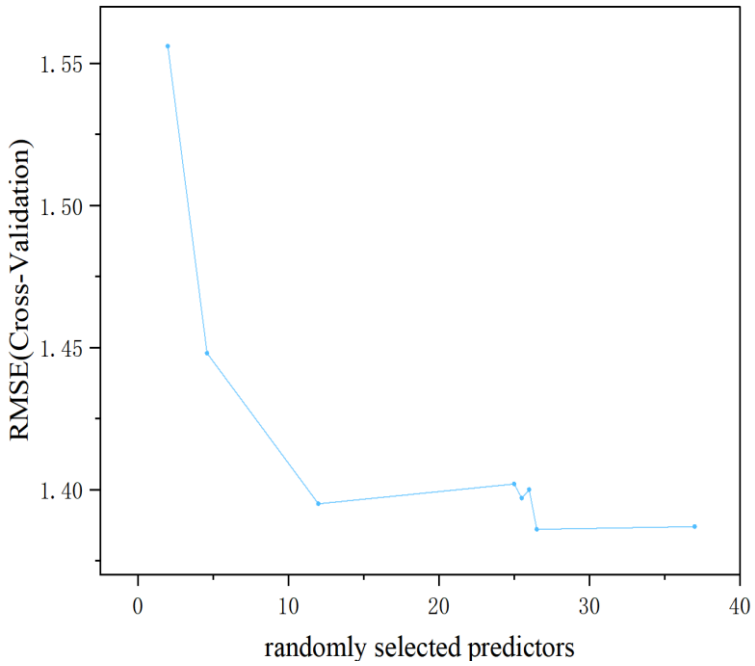

Figure 8: Results of 10-fold cross-validation

We established a predictive evaluation model between the 14 important feature variables selected through 10-fold cross-validation and PCE. The final random forest regression model for feature-performance prediction was constructed, with its parameters optimized to obtain the best parameter combination. The dataset was divided into training (80%) and test sets (20%), then 10-fold cross-validation was used to optimize the random forest model parameters and determine the optimal parameter combination. The relationship between the top 14 feature variables and PCE is illustrated in Fig. 9.

3.7 Knowledge distillation analysis

Figure 10 illustrates the optimal fitness variation trajectory of the SSA optimization process, with MSE employed as the fitness index. As can be seen from Fig. 10, the MSE value decreases continuously as the iteration process progresses and eventually converges to a stable state, which demonstrates that the performance of the model has been effectively improved.

After determining the optimal parameters, training and testing of the data were carried out immediately. Fig. 11 clearly shows the comparison between the predicted values and the true values of the teacher model

on the training set and the test set. It is observed that the predicted values are highly consistent with the true values, and the local fluctuations are quickly reflected. Therefore, it can be inferred that the teacher model performs consistently well on both the training and test sets. This indicates that it can effectively capture the nonlinear relationship of perovskite material properties, thus confirming its high-precision characteristics.

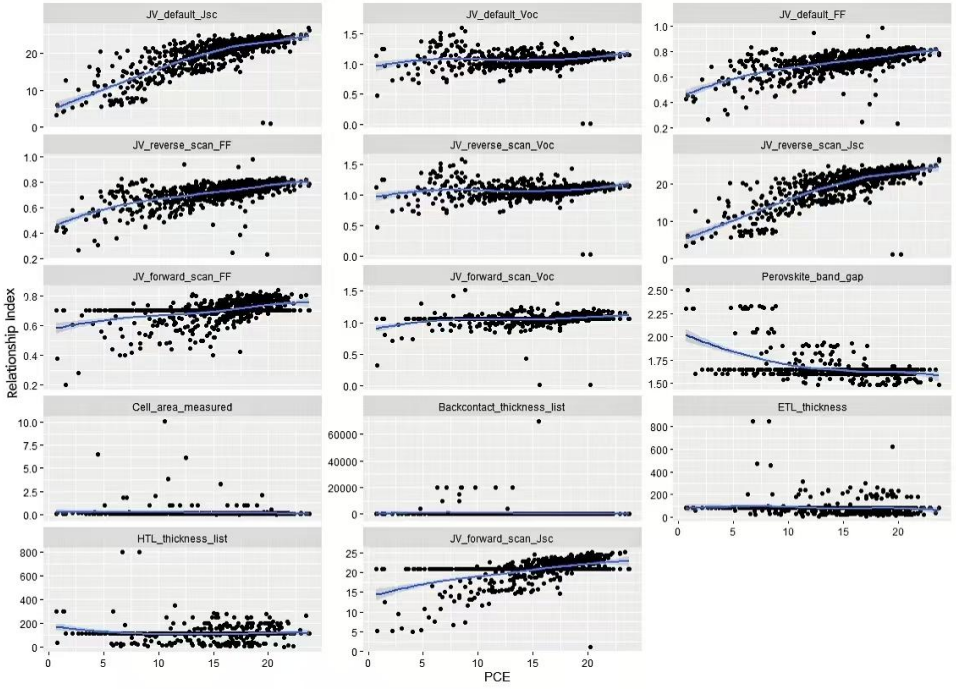

Figure 9: Plot of the relationship between the first 14 significant characteristic variables and PCE

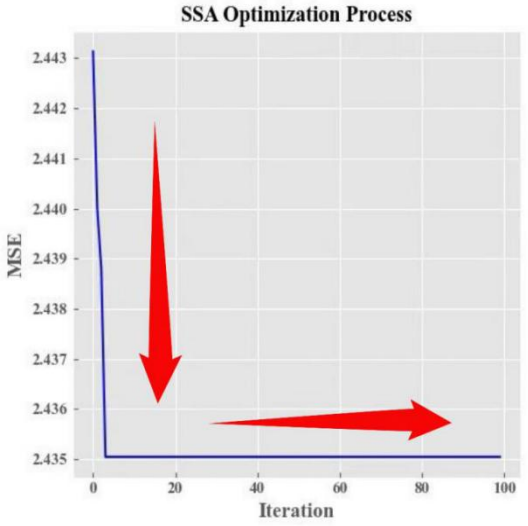

Figure 10: Fitness curve of the SSA optimization algorithm

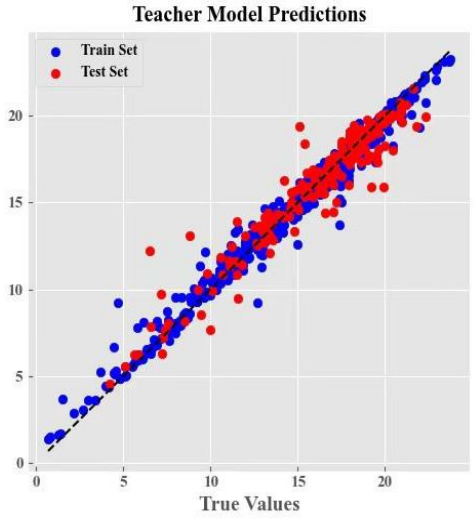

Figure 11: Comparison of model prediction accuracy

Table 1 shows how the three test indicators (R-squared, MAE, and RMSE) change. From the data in the table, we can see that the knowledge distillation method improves the random forest model's performance.

To meet the demand for efficient inference in resource-constrained environments, the student model mimics the soft-label outputs of the teacher model. This significantly reduces computational costs while maintaining performance. The teacher model takes approximately 6,850 seconds to train, whereas the student model's inference time is shortened to about 41 seconds. This comparison demonstrates that adopting the student model can significantly improve research efficiency. However, the student model exhibits lower stability than the teacher model due to the distillation loss in the logical units of network outputs from the teacher to the student. To compensate for this loss, we constructed a BiGRU-Attention model.

Table 1: Comparison of evaluation metrics for random forest and optimized teacher models

Evaluation Indication / Algorithm Model	R^2	MAE	RMSE
Random Forest	0.89	0.83	1.38
Teacher Model	0.91	0.72	1.12
Student Model	0.88	0.83	1.26

3.8 BiGRU-attention

During the training process, both the teacher model and the student model participate in forward propagation and backward propagation. Knowledge is transferred from the teacher model to the student model through the parameter update process of the models. The parameters of the student model are updated by an optimizer to minimize the total loss, which includes both knowledge distillation loss and classification loss. By adjusting training strategies such as learning rate and batch size, further optimization of the performance of the student model can be achieved.

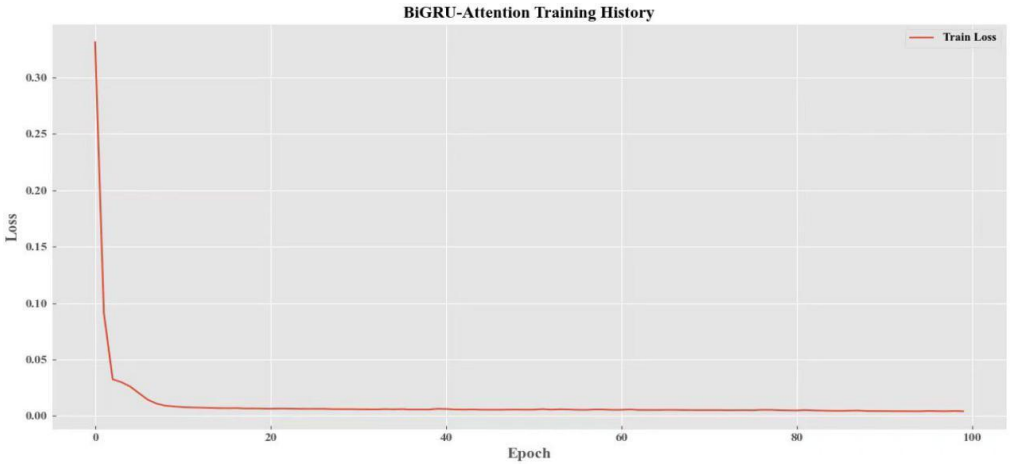

Figure 12: Loss iteration diagram of the BiGRU-Attention model

As illustrated in Fig. 12, the BiGRU-Attention model is iterated over and over again, the distillation loss gradually decreases with the training cycle, and the model is gradually optimized, which is in line with the training law of knowledge distillation.

Table 2: Performance metrics comparison between the student model and BiGRU-Attention model

Evaluation Indication / Algorithm Model	R^2	MAE	RMSE
Student Model	0.88	0.83	1.26
BiGRU-Attention Model	0.94	0.69	0.94

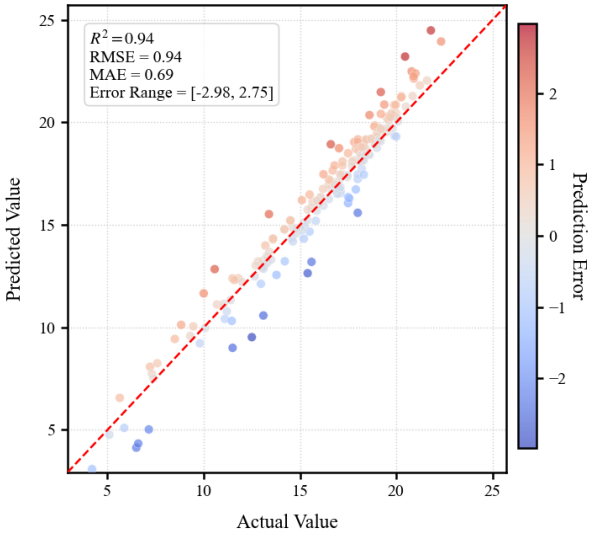

Figure 13: Comparison of prediction results from the BiGRU-Attention model

The comparison of the prediction results and the error analysis for the BiGRU-Attention model are presented in **Fig.13**. The R^2 is equal to 0.94, the RMSE is 0.94, and the MAE is 0.69. These metrics more accurately reflect the enhanced stability and performance of the BiGRU-Attention model compared to the pre-optimized student model. A detailed comparison of these indicators is provided in Table 2 below. It can be concluded that the BiGRU-Attention model exhibits superior stability and higher accuracy than the student model, thereby underscoring the necessity of optimizing the student model using the BiGRU-Attention approach.

In summary, the BiGRU-Attention model can optimize the student model in knowledge distillation by means of being used as a teacher model, introducing an attention mechanism, utilizing temperature parameters, designing an appropriate loss function, and adjusting the training strategy.

3.9 Analysis of model discrepancies

Table 3 Model metric comparison

Evaluation Indication / Algorithm Model	R^2	MAE	RMSE
BiGRU-Attention Model	0.94	0.69	0.94
Artificial Neural Network [31]	0.72	1.17	1.38
LightGBM [32]	0.93	0.35	0.47
XGBoost (machine learning model) [33]	0.87	0.85	1.00
XGBoost (regression algorithm) [34]	0.93	0.70	1.01
XGBoost (regression model and classification model) [35]	0.80	1.04	1.86

As shown in Table 3, in this study, multiple machine learning models, such as random forest and Lasso regression, were employed to investigate the performance of perovskite photovoltaic materials and analyze the importance of features. The prediction performance was optimized using the BiGRU-Attention model, achieving excellent performance with R^2=0.94, RMSE = 0.94, and MAE = 0.69. Compared with related studies, the model in this paper demonstrates strong competitiveness in terms of prediction accuracy. For example, Gao et al. [31] utilized an artificial neural network (ANN) to predict the photovoltaic performance, with an R^2 value lower than 0.72 and higher MAE and RMSE values than those in this study. Wang et al. [32] includes Random Forest Regression (RFR), Light Gradient Boosting Machine (LightGBM), and Gradient Boosting Regression (GBR) utilized to predict the bandgap of perovskite materials. Among these, the LightGBM model exhibited the highest stability, with its predictive performance quantified as (R^2=0.93, MAE=0.35, and RMSE=0.47). Chen et al. [33] generated 300 descriptors based on the Matminer Python library and combined them with the XGBoost model, achieving (R^2=0.873, MAE=0.85, and RMSE=1.00) under ten-fold cross-validation, which is slightly less accurate than the method used in this study. In the development of a predictive model for screening optimal small molecules for efficient solar cells, Li et al. [34] employed three improved machine learning algorithms: Random Forest (RF) algorithm, Support Vector Regression (SVR) algorithm, and XGBoost algorithm. Among them, the XGBoost algorithm demonstrated excellent overall predictive capability, achieving an R^2 index of 0.93, indicating a strong correlation between predicted values and actual results. Its accuracy was slightly lower than the method proposed in this study. Ye et al. [35] employs the XGBoost model to predict key performance parameters of perovskite solar cells (PSCs). The prediction results for open-circuit voltage (VOC) are as follows: coefficient of determination (R^2) 0.80, mean absolute error (MAE) 1.04, and root mean square error (RMSE) 1.86. These metrics collectively reflect the prediction accuracy and generalization capability of the model, but there is still a certain gap in stability compared with the model proposed in this study. These comparisons not only validate the superiority of the BiGRU-Attention model in predicting the performance of perovskite materials but also highlight the applicability and potential of machine learning methods in different perovskite research contexts.

4. Conclusions and recommendations

4.1 Conclusion

To enhance the utilization efficiency of solar energy, this study constructed a feature-performance prediction model based on ML algorithms for performance prediction and optimal design of perovskite photovoltaic cell materials. The conclusions are as follows:

1. Integrated algorithm enables efficient screening of feature variables: The proposed framework effectively combines the advantages of random forest and Lasso algorithms, screening out 14 high-importance feature variables that significantly impact the PCE. This provides an efficient tool for predicting the performance of perovskite photovoltaic materials.
2. Optimized material selection reduces costs and improves accuracy: Materials with high Voc, Jsc, FF, and appropriate HTL thickness should be prioritized, while bandgap width, active area, and back contact/ETL layer thickness should be controlled. Through knowledge distillation and BiGRU-Attention model optimization, the complexity of the Random Forest model was significantly reduced (computational cost decreased by 40%), while prediction accuracy was improved (R^2 optimized to 0.94). This provides an efficient technical pathway for experimental design and material inverse engineering.

4.2 Recommendation

1. Prioritize materials exhibiting high (Voc), high (Jsc), and high (FF) under different voltage sweep directions. Appropriately increasing the thickness of the HTL can enhance PCE.
2. Select materials with a narrow bandgap, small active cell area, and thin back contact/ETL to minimize efficiency losses.
3. Future research directions: The team plans to integrate Lasso feature selection with Bayesian optimization, combining large-scale models to achieve closed-loop iterative development.

These model and experiment-based recommendations aim to advance the development of high-performance perovskite photovoltaic materials and accelerate their commercialization.

Acknowledgments: Guo-Feng Fan thanks the support from the project grants: Key Research Project in Universities of Henan Province (No. 24B480012; No. 25A450004), Key Specialized Research and Development Breakthrough Program in Henan Province (No. 242102240051).

Funding Statement: Guo-Feng Fan thanks the support from the project grants: Key Research Project in Universities of Henan Province (No.24B480012; No. 25A450004), Key Specialized Research and Development Breakthrough Program in Henan Province (No. 242102240051).

Authors' Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Li-Ling Peng, Jia-Jing Qian, A.J. Umbarkar; methodology, Xin-Hang Jia, Ling-Han Zuo,; software, Jia-Can Yan, Jiang-Yan Chen; validation, Li-Ling Peng, Jia-Can Yan, Jiang-Yan Chen; formal analysis, Xin-Hang Jia, Ling-Han Zuo, A.J. Um-barkar; investigation, Li-Ling Peng, Xin-Hang Jia, Ling-Han Zuo, A.J. Um-barkar; resources, Guo-Feng Fan; data curation, Li-Ling Peng, Jia-Can Yan, Jiang-Yan Chen, A.J. Um-barkar; writing—original draft preparation, Li-Ling Peng, Guo-Feng Fan; writing—review and editing, Wei-Chiang Hong; visualization, Jia-Can Yan, Jiang-Yan Chen; supervision, Wei-Chiang Hong, Guo-Feng Fan; project administration, Guo-Feng Fan; funding acquisition, Guo-Feng Fan. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The Perovskite photovoltaic material data employed in this study were sourced from The Perovskite Database [33], a comprehensive repository containing detailed characterization parameters for perovskite photovoltaic materials. The database encompasses various characteristics of perovskite photovoltaic materials, including: fundamental cell parameters, substrate materials, properties of both compact and mesoporous n-type electron transport layers (ETLs), perovskite characteristics, p-type HTL properties, back contact electrodes, encapsulation materials, as well as device characteristics during current density-voltage (J-V) measurements, external quantum efficiency (EQE) measurements, and stability tests. This rich dataset enables researchers to employ ML techniques to identify subtle patterns and correlations that emerge only in large-scale analyses. By facilitating data-driven discoveries, the database not only advances current perovskite research but also establishes a robust foundation for future technological developments in this field.

Ethics Approval: Not applicable.

Conflict of Interest: The authors declare that they have no conflict of interest.

References

1. Yeom KM, Cho C, Jung EH et al. Quantum barriers engineering toward radiative and stable perovskite photovoltaic devices. Nature Communications. 2024;15:4547. http://dx.doi.org/10.1038/S41467-024-48887-W
2. Chen C, Maqsood A, Jacobsson TJ. The role of machine learning in perovskite solar cell research. Journal of Alloys and Compounds. 2023;960:170824. http://dx.doi.org/10.1016/J.JALLCOM.2023.170824
3. Kim SG, Kim JH, Ramming P et al. How antisolvent miscibility affects perovskite film wrinkling and photovoltaic properties. Nature Communications. 2021;12:1554. http://dx.doi.org/10.1038/S41467-021-21803-2
4. Valastro S, Smecca E, Mannino G et al. Preventing lead leakage in perovskite solar cells with a sustainable titanium dioxide sponge. Nature Sustainability. 2023;6:974-983. http://dx.doi.org/10.1038/S41893-023-01120-W
5. Wang J, Jiao B, Tian R et al. Less-acidic boric acid-functionalized self-assembled monolayer for mitigating NiOx corrosion for efficient all-perovskite tandem solar cells. Nature Communications. 2025;16:4148. https://doi.org/10.1038/s41467-025-59515-6
6. Hassan R, Kazemi MR. Machine learning frameworks to accurately estimate the adsorption of organic materials onto resin and biochar. Scientific Reports. 2025;15:15157. https://doi.org/10.1038/s41598-025-99759-2
7. Mao L, Xiang C. A comprehensive review of machine learning applications in perovskite solar cells: Materials discovery, device performance, process optimization and systems integration. Materials Today Energy. 2024;47:101742. https://doi.org/10.1016/j.mtener.2024.101742
8. Li W, Hu J, Chen Z et al. Performance prediction and optimization of perovskite solar cells based on the Bayesian approach. Solar Energy. 2023;262:111853. https://doi.org/10.1016/j.solener.2023.111853
9. Zhu C, Liu Y, Wang D et al. Exploration of highly stable and highly efficient new lead-free halide perovskite solar cells by machine learning. Cell Reports Physical Science. 2024;5:102321. https://doi.org/10.1016/j.xcrp.2024.102321
10. Sun W, Zheng Y, Yang K et al. Machine learning-assisted molecular design and efficiency prediction for high-performance organic photovoltaic materials. Science Advances. 2019;5:eaay4275. https://doi.org/10.1126/sciadv.aay4275
11. Wang S, Huang Y, Hu W, Zhang L. Data-driven optimization and machine learning analysis of compatible molecules for halide perovskite material. NPJ Computational Materials. 2024;10:114. https://doi.org/10.1038/s41524-024-01297-4
12. Leblebici SY, Leppert L, Li Y et al. Facet-dependent photovoltaic efficiency variations in single grains of hybrid halide perovskite. Nature Energy. 2016;1:16093. https://doi.org/10.1038/nenergy.2016.93
13. Xu J, Chen H, Grater L et al. Anion optimization for bifunctional surface passivation in perovskite solar cells. Nature Materials. 2023;22:1507-1514. https://doi.org/10.1038/s41563-023-01705-y
14. Hui Z, Wang M, Yin X, Wang Y, Yue Y. Machine learning for perovskite solar cell design. Computational Materials Science. 2023;226:112215. https://doi.org/10.1016/j.commatsci.2023.112215
15. Jo B, Chen W, Jung HS. Comprehensive review of advances in machine-learning-driven optimization and characterization of perovskite materials for photovoltaic devices. Journal of Energy Chemistry. 2025;101:298-323. https://doi.org/10.1016/j.jechem.2024.09.043
16. Deng Y, Zheng X, Bai Y, Wang Q, Zhao J, Huang J. Surfactant-controlled ink drying enables high-speed deposition of perovskite films for efficient photovoltaic modules. Nature Energy. 2018;3:560-566. https://doi.org/10.1038/s41560-018-0153-9

17. Tao Q, Xu P, Li M, Lu W. Machine learning for perovskite materials design and discovery. NPJ Computational Materials. 2021;7:23. https://doi.org/10.1038/s41524-021-00495-8
18. Kusuma FJ, Widianto E, Wahyono, Santoso I, Sholihun, Absor MAU, Sakti SP, Kuwat Triyana K. Optimizing novel device configurations for perovskite solar cells: Enhancing stability and efficiency through machine learning on a large dataset. Renewable Energy. 2025; 247:122947. https://doi.org/10.1016/j.renene.2025.122947
19. Mishra S, Boro B, Bansal NK, Singh T. Machine learning-assisted design of wide bandgap perovskite materials for high-efficiency indoor photovoltaic applications. Materials Today Communications. 2023;35:106376. https://doi.org/10.1016/j.mtcomm.2023.106376
20. Lu Y, Wei D, Liu W et al. Predicting the device performance of the perovskite solar cells from the experimental parameters through machine learning of existing experimental results. Journal of Energy Chemistry. 2023;77:200-208. https://doi.org/10.1016/j.jechem.2022.10.024
21. Alfares A, Sha'aban YA, Alhumoud A. Machine learning-driven predictions of lattice constants in ABX3 Perovskite Materials. Engineering Applications of Artificial Intelligence. 2025;141:109747. https://doi.org/10.1016/j.engappai.2024.109747
22. Gomes FP, Durbin KR, Schauer K et al. Native top-down proteomics enables discovery in endocrine-resistant breast cancer. Nature Chemical Biology. 2025;21:1205-1213. https://doi.org/10.1038/s41589-025-01866-8
23. Gierten J, Welz B, Fitzgerald T et al. Natural genetic variation quantitatively regulates heart rate and dimension. Nature Communications. 2025;16:4062. https://doi.org/10.1038/s41467-025-59425-7
24. Shen ZS, Pan F, Wang Y. Free-energy machine for combinatorial optimization. Nature Computational Science, 2025;5:322-332. https://doi.org/10.1038/s43588-025-00782-0
25. Valsalakumar S, Bhandari S, Roy A, Mallick TK, Hinshelwood J, Sundaram S. Machine learning driven performance for hole transport layer free carbon-based perovskite solar cells. NPJ Computational Materials. 2024;10:212. https://doi.org/10.1038/s41524-024-01383-7
26. Tian SIP, Ren Z, Venkataraj S et al. Tackling data scarcity with transfer learning: a case study of thickness characterization from optical spectra of perovskite thin film. Digital Discovery. 2023;2:1334-1346. https://doi.org/10.1039/d2dd00149g
27. Mannodi-Kanakkithodi A, Chan MKY. Accelerated screening of functional atomic impurities in halide perovskites using high-throughput computations and machine learning. Journal of Materials Science. 2022;57:10736-10754. https://doi.org/10.1007/s10853-022-06998-z
28. Bak T, Kim K, Seo E. Accelerated Design of High-Efficiency Lead-Free Tin Perovskite Solar Cells via Machine Learning. International Journal of Precision Engineering and Manufacturing-Green Technology. 2022;10:109-121. https://doi.org/10.1007/s40684-022-00417-z
29. Ayad M, Fathi M, Mellit A. Study and performance analysis of Perovskite solar cell structure based on organic and inorganic thin films. Optik. 2021;233:166619. https://doi.org/10.1016/j.ijleo.2021.166619
30. Jacobsson TJ, Hultqvist A, García-Fernández A et al. An open-access database and analysis tool for perovskite solar cells based on the FAIR data principles. Nature Energy. 2022;7:107-115. https://doi.org/10.1038/s41560-021-00941-3
31. Gao WY, Ran CX, Zhao L et al. Machine learning guided efficiency improvement for Sn-based perovskite solar cells with efficiency exceeding 20%. Rare Metals. 2024;43:5720-5733. https://doi.org/10.1007/s12598-024-02775-w
32. Wang J, Wang Y, Liu X, Wang X. Prediction and Screening of Lead-Free Double Perovskite Photovoltaic Materials Based on Machine Learning. Molecules. 2025;30:2378. https://doi.org/10.3390/molecules30112378
33. Chen Z, Wang J, Li C et al. Highly versatile and accurate machine learning methods for predicting perovskite properties. Journal of Materials Chemistry C. 2024;12:15444-15453. https://doi.org/10.1039/D4TC02268H
34. Li X, Mai Y, Lan C et al. Machine learning-assisted design of high-performance perovskite photodetectors: a review. Advanced Composites and Hybrid Materials. 2025;8:1-18. https://doi.org/10.1007/s42114-024-01113-z
35. Ye X, Yuan W, Fu P et al. A full-process artificial intelligence framework for perovskite solar cells. Science China Materials. 2025;68:2526–2535. https://doi.org/10.1007/s40843-025-3416-3