



Prediction of unmeasured optoelectronic properties of Au/(GO:P₃C₄MT)/SiO₂ heterojunctions via ensemble machine learning models

Ömer Arslan¹, İhsan Tuğal^{2,*}, Halil Seymen³, and Şükrü Karataş⁴

¹ Faculty of Engineering and Architecture, Department of Mechanical Engineering, Muş Alparslan University, 49250 Muş, Turkey

² Faculty of Engineering and Architecture, Department of Software Engineering, Muş Alparslan University, 49250 Muş, Turkey

³ Technical Sciences Vocational High School, Department of Electronics and Automation, Muş Alparslan University, 49250 Muş, Turkey

⁴ Faculty of Sciences, Department of Physics, Kahramanmaraş Sütçü İmam University, 46100 Kahramanmaraş, Turkey

Received: 22 December 2025

Accepted: 5 May 2026

Published online:
12 May 2026

© The Author(s), 2026

ABSTRACT

This study presents a statistically validated machine learning approach for predicting unmeasured optoelectronic responses at an Au/SiO₂/n-Si semiconductor interface incorporating a graphene oxide (GO) and poly(3-cyclohexyl-4-methyl-2,5-thiophene) (P₃C₄MT) hybrid interfacial layer. Although GO–polymer hybrid interfaces exhibit promising photoactive properties, comprehensive experimental characterization of current–light intensity relationships is often constrained by the high cost and time demands associated with high-illumination measurements. The experimental dataset used for model training was obtained by measuring current–voltage characteristics at room temperature, with voltage varied from −2 to +2 V in 0.02 V increments under illumination intensities ranging from 20 to 100 mW/cm² at 20 mW/cm² intervals. Five ensemble and tree-based supervised regression models (Random Forest, Gradient Boosting, LightGBM, HistGradientBoosting, and ExtraTrees) were employed using voltage and light intensity as input features to predict current responses at intermediate and extended illumination levels (10, 30, 50, 70, 90, and 110 mW/cm²). Model performance was evaluated using R², mean squared error (MSE), mean absolute error (MAE), and mean absolute percentage error (MAPE), with all models achieving high predictive accuracy (R² > 0.99). Among them, the ExtraTrees regression model consistently produced the lowest prediction errors across all illumination conditions. To ensure statistical rigor, a two-way ANOVA was conducted to examine the effects of model type and illumination intensity, revealing statistically significant main and interaction effects ($p < 0.001$). Post hoc Tukey HSD analysis further confirmed the statistically superior performance of the ExtraTrees model. These results demonstrate that reliable interpolation and controlled extrapolation of optoelectronic behavior can be achieved using sparse experimental data, providing a cost-effective, scalable,

Address correspondence to E-mail: i.tugal@alparslan.edu.tr

and statistically robust methodology for the characterization and design of hybrid semiconductor interfaces in photonic, photovoltaic, and photosensor applications.

1 Introduction

The global energy demand has experienced a consistent increase, driven by rising living standards, population growth, and rapid technological advancements. The depletion of finite fossil fuel resources, coupled with their adverse environmental impacts, has accelerated the transition toward renewable energy sources and catalyzed a surge in research in this domain worldwide [1, 2]. Among the various renewable energy technologies, photovoltaic and photodiode materials, which convert solar energy into electrical energy, have drawn considerable attention.

The fabrication of advanced devices is fundamentally based on semiconductors, which include both elemental semiconductors such as silicon (Si) and germanium (Ge), as well as compound semiconductors like gallium arsenide (GaAs). In addition, insulating oxide materials such as silicon dioxide (SiO_2) play a critical role as dielectric layers in metal–oxide–semiconductor (MOS) architectures. To enhance device performance, various interfacial materials have been extensively investigated, including polymer- or graphene oxide (GO)-based interfacial layers integrated with semiconductor substrates, as well as MOS-structured Schottky diodes [3]. The sustained interest in MOS-based devices stems from their simplified fabrication protocols and cost-effectiveness [4]. Graphene oxide, the functionalized derivative of graphene, is particularly distinguished by its tunable bandgap, large surface area, and rich oxygen-containing functional groups. Similarly, conjugated polymers exhibit significant conductivity when doped or integrated into hybrid systems. Recent research indicates that the integration of GO with the polymer poly(3-cyclohexyl-4-methyl-2,5-thiophene) ($\text{P}_3\text{C}_4\text{MT}$) significantly enhances optoelectronic properties, positioning it as a promising candidate for utilization in photovoltaic devices, photodiodes, and organic light-emitting diodes (OLEDs) [5].

Advancing materials for energy conversion systems, particularly regarding their optoelectronic performance and interfacial charge dynamics, remains a

critical engineering challenge. Characterizing these hybrid materials under varying environmental and illumination conditions often necessitates time-intensive and costly experimental procedures. For instance, evaluating photovoltaic and photosensing capacities across a broad range of light intensities requires extensive datasets, which can be both economically demanding and temporally inefficient. To mitigate these limitations, machine learning (ML)-based modeling approaches have emerged as a powerful alternative. By leveraging a curated set of experimental data, ML techniques can effectively capture material behaviors, thereby reducing the necessity for exhaustive repetitive measurements and providing a robust computational framework for future experimental designs. [6–8].

In this study, the optoelectronic response of a graphene oxide–polymer-based hybrid interface was investigated using a combination of experimental measurements and machine learning modeling. Experimental current–voltage data obtained under selected illumination conditions were used to train predictive models capable of estimating device behavior at unmeasured light intensities. This approach aims to extend the characterization range while reducing the need for exhaustive experimental measurements. The primary objective is to establish a statistically reliable framework for evaluating semiconductor interfaces using limited data without compromising predictive accuracy.

Recent scholarly literature highlights the practical application of machine learning (ML) techniques in predicting the electrical and electrochemical properties of graphene and its derivatives within electronic structures. Yogesh et al. (2025) attained accurate predictions of the specific capacitance of graphene oxide (GON)-based nanomaterials through the application of supervised ML algorithms, including artificial neural networks (ANN), regression trees (RT), support vector machines (SVM), and linear regression (LR), drawing upon data aggregated from 64 investigations [9]. Similarly, Barkhordari et al. (2023) employed Gaussian process regression (GPR), kernel

ridge regression (KRR), support vector regression (SVR), and ANN algorithms to forecast the electrical parameters of Schottky diodes integrated with graphene-doped nanocomposite interlayers, determining that SVR emerged as the most effective model [10]. In a related endeavor, Ersöz et al. (2025) developed models to characterize the temperature-dependent current–voltage (I – V) behavior of nanodiodes incorporating lanthanum-doped graphene quantum dots, utilizing K-nearest neighbors (KNN), decision trees, and gradient boosting, with gradient boosting yielding the highest level of accuracy [11]. Additionally, Ocaya et al. (2023) successfully predicted the intrinsic parameters of graphene oxide (GO)-doped Schottky photo-diodes under illuminated conditions with the use of ANN, thereby obviating the necessity for thermionic emission expressions [12]. Furthermore, García-Córdova et al. (2024) investigated the nonlinear optical properties of graphene. They demonstrated that decision tree-based ML algorithms, such as Random Forest (RF) and gradient boosted regression (GBR), surpassed their linear counterparts in performance [13]. Lastly, Rabehi et al. (2020) conducted a comparative analysis of swarm intelligence algorithms for predicting parameters of Schottky diodes, identifying the ant lion optimizer (ALO) algorithm as the most accurate, as it produced the lowest root mean square error (RMSE) value [14].

This study provides a systematic and statistically grounded evaluation of ensemble-based machine learning models for predicting the optoelectronic behavior of polymer-based hybrid interfaces. By comparing five advanced algorithms, the proposed approach effectively captures the current–light intensity relationships that are difficult to represent using conventional curve-fitting techniques.

The results demonstrate that reliable predictions can be achieved even with limited experimental data, highlighting the potential to reduce extensive measurement requirements. Particular attention is given to model performance under low-illumination conditions, which are critical for practical photosensor applications yet often underexplored.

Furthermore, the integration of rigorous statistical validation, including two-way ANOVA and Tukey HSD analysis, ensures that performance differences between models are both quantitatively and statistically significant. Beyond the specific GO:P₃C₄MT system, the proposed framework is inherently extensible and can be adapted to a wide range of optoelectronic

materials, supporting more efficient and data-driven experimental design.

The remainder of this paper is organized as follows: Section 2 details the experimental methodology and material preparation, including the fabrication of the GO–P₃C₄MT interface and the protocols for I – V measurements under varying light intensities. This section also outlines the dataset, feature selection, and the theoretical basis of the five supervised regression algorithms employed: Random Forest, Gradient Boosting, LightGBM, HistGradientBoosting, and ExtraTrees. In addition, the section describes the performance evaluation metrics (R^2 , MSE, MAE, and MAPE) used to ensure predictive reliability and statistical validation and ANOVA-based error analysis. Section 3 presents the Results and Discussion, providing a comparative analysis of model predictions against experimental data, cross-validation outcomes, and performance at both intermediate and extrapolated illumination levels. Finally, Section 4 concludes the study by synthesizing key findings and discussing the broader implications of ML-assisted characterization for optoelectronic materials.

2 Methods

2.1 Material preparation and experimental design

A hybrid interfacial layer was fabricated using a 1:1 weight ratio of graphene oxide (GO) and poly(3-cyclohexyl-4-methyl-2,5-thiophene) (P₃C₄MT), following the protocol established in [5]. The structural configuration of the material is illustrated schematically in Fig. 1, which shows the layered architecture consisting of the Au top contact, the GO:P₃C₄MT organic interfacial layer, the SiO₂ insulating layer, and the n-type silicon substrate with an Au back contact. The fabrication of the Au/(GO:P₃C₄MT)/SiO₂/n-Si/Au heterostructure was carried out through a series of precise chemical and physical deposition steps. Initially, n-type Si(100) wafers (400 μm thick, 1–10 $\Omega\cdot\text{cm}$ resistivity) were cleaned using the standard RCA procedure (involving acetone, propanol, NH₄OH, H₂O₂, and HCl) to eliminate organic contaminants and the native oxide layer. Following cleaning, an ohmic back contact was formed by thermally evaporating high-purity gold (Au, 99.99%) at a vacuum pressure of 2×10^{-6} Torr,

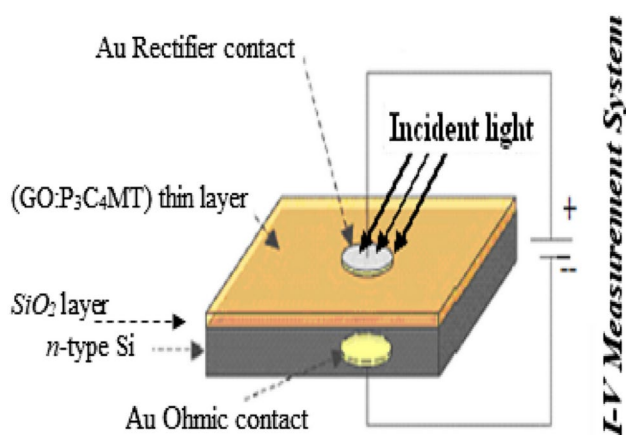


Fig. 1 Cross-section view of the Au/(GO:P3C4MT)/SiO₂/n-Si semiconductor structure [5]

followed by annealing at 550 °C for 3 min under a nitrogen atmosphere to prevent oxidation and ensure low contact resistance.

Subsequently, a SiO₂ insulating layer was deposited onto the front surface of the n-Si substrate via a modified thermal evaporation technique at 150 °C and 10⁻³ Torr, utilizing SiH₄ and O₂ as the precursor

sources. For the interfacial layer, a hybrid solution of graphene oxide (GO) and P₃C₄MT (1:1 molar ratio, 10⁻³ mol L⁻¹) was prepared in chloroform. This solution was then deposited onto the SiO₂/n-Si structure by spin-coating at 1500 rpm for 30 seconds. Finally, the device was completed by thermally evaporating circular Au rectifying contacts (1 mm diameter, 7.85 × 10⁻³ cm²) through a shadow mask at 10⁻⁶ Torr. The final schematic of the fabricated MOS-type Schottky structure is illustrated in Fig. 1.

The electrical performance of the fabricated Au/(GO:P₃C₄MT)/SiO₂/n-Si structure was evaluated through current–voltage (*I*–*V*) measurements carried out at room temperature under both dark conditions and various illumination intensities. The experimental setup used for these measurements is presented in Fig. 2, which includes a solar simulator as the light source and a precision source meter for electrical characterization. In addition to electrical measurements, the surface morphology of the GO:P₃C₄MT organic semiconductor was analyzed using atomic force microscopy (AFM) to assess film uniformity and surface features. The prepared GO–polymer hybrid film was subjected to photon excitation at fixed voltage intervals for experimental assessment. Photocurrent

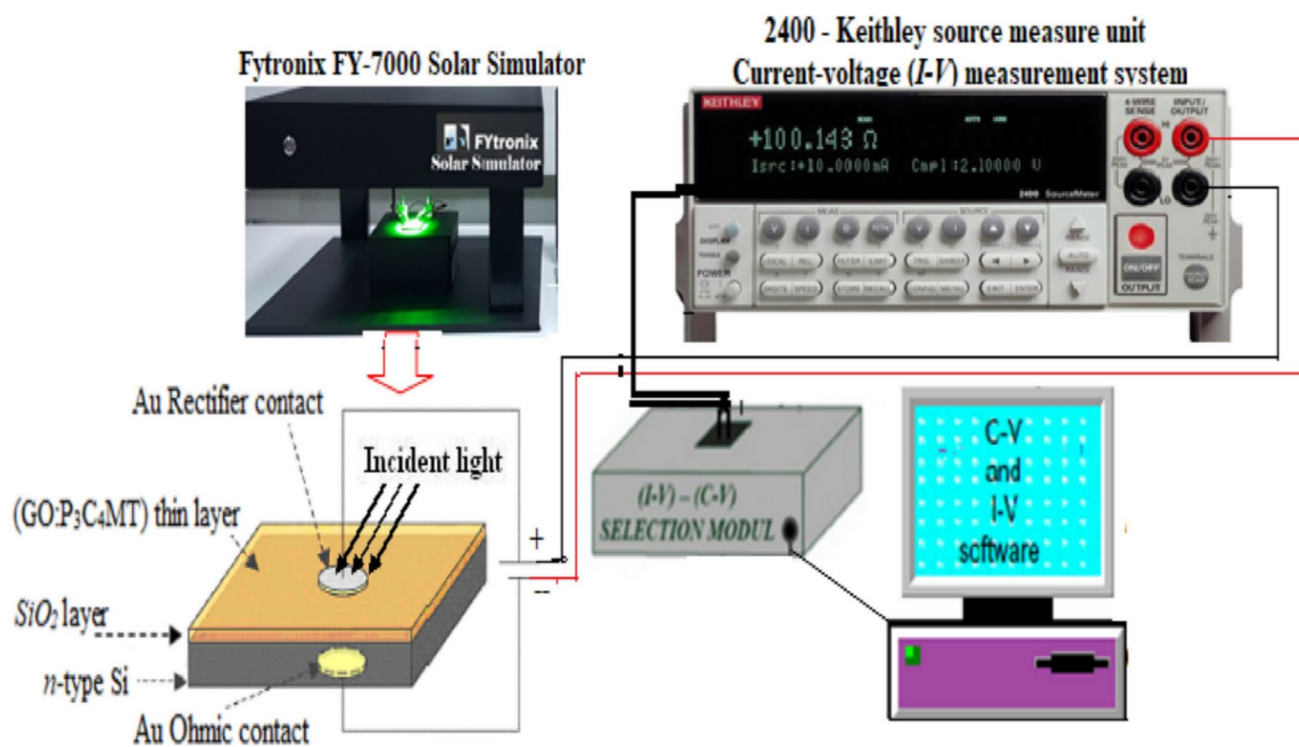


Fig. 2 *I*–*V* measurement system of the Au/(GO:P3C4MT)/SiO₂/n-Si semiconductor structure to collect experimental data [5]

measurements were performed using a Keithley 2400 source meter at room temperature. The experimental design included current–voltage (I – V) measurements at five distinct illumination intensities: 20, 40, 60, 80, and 100 mW/cm². For each illumination level, the current was recorded across a voltage range of -2.0 V to $+2.0$ V, with increments of 0.02 V, facilitating the construction of a comprehensive dataset.

This dataset provides sufficient resolution to understand the complex electrical response of the hybrid GO–polymer interface and serves as the basis for machine learning modeling.

2.2 Dataset and features

The experimental data gathered consisted of three primary components: “Voltage” (independent variable), “Light” (illumination intensity, independent variable), and “Current” (dependent variable). For machine learning purposes, Voltage and Light were used as input features, while Current was set as the target variable. The ‘Light’ variable was recognized as a critical parameter, enabling the model to accurately represent the effect of illumination intensity across both training and testing phases.

Device hysteresis was not explicitly modeled in this study. The dataset consists of steady-state measurements, and therefore the models learn a single-valued mapping between inputs and outputs. Incorporating hysteresis would require time-dependent or cyclic measurement data.

For each illumination level, the current was measured over a voltage range from -2.0 V to $+2.0$ V in 0.02 V increments, resulting in 201 data points per illumination intensity, and a total of 1005 training measurements. For the purpose of model training, only experimental measurements corresponding to illumination levels of 20, 40, 60, 80, and 100 mW/cm² were employed. During the testing phase, the model was tasked with predicting the system’s response for unmeasured intermediate intensities (30, 50, 70, 90 mW/cm²) as well as extrapolated boundary conditions (10 and 110 mW/cm²) that were not present in the training dataset.

Feature importance analysis indicates that voltage is the dominant predictor of current due to its direct role in defining I – V characteristics. However, light intensity also contributes significantly by modulating

the magnitude of the current. The combined effect of both features is essential for accurate modeling.

No outlier removal or smoothing was applied, in order to preserve the raw physical characteristics and inherent experimental noise. Furthermore, as tree-based ensemble algorithms are inherently invariant to feature scaling, no normalization or standardization was applied to the inputs. This dataset provides sufficient resolution to characterize the relationship between applied voltage, illumination intensity, and current, enabling reliable predictions at intermediate light intensities using supervised regression models.

The results demonstrate that high predictive accuracy can be achieved with a limited dataset. However, the minimum data requirement depends on the variability of the system. In general, maintaining sufficient coverage of the voltage range is more critical than increasing the number of illumination levels. In addition, the ensemble-based methods employed in this study do not require large-scale training data, unlike more complex deep learning models, and can achieve robust performance even with relatively small experimental datasets.

2.3 Machine learning models

In this study, five supervised regression algorithms were implemented, including Random Forest, Gradient Boosting, LightGBM, HistGradientBoosting, and ExtraTrees. Although all these models are tree-based, they differ in their learning strategies and structures, falling broadly into Bagging (Random Forest, ExtraTrees) and Boosting (Gradient Boosting, LightGBM, HistGradientBoosting) categories. These models were selected due to their robust capability to represent intricate dependencies, which are frequently encountered in photovoltaic systems. In contrast to linear models, tree-based ensembles effectively manage complex interactions among features, such as light intensity and voltage.

Furthermore, these models are particularly well-suited for small to medium-sized experimental datasets, where deep learning techniques may not be feasible. Specifically, Random Forest and ExtraTrees, through parallel training of independent trees (Bagging), prioritize the reduction of variance by employing ensemble averaging and randomness. In contrast, boosting algorithms (Gradient Boosting, LightGBM, and HistGradientBoosting) are trained sequentially, iteratively minimizing residual errors from the

previous stage to reduce bias and thus enhance prediction accuracy under varying light conditions. Among the Boosting methods, LightGBM and HistGradient-Boosting were chosen due to their superior computational efficiency and scalability, utilizing histogram-based approaches for faster training times compared to traditional Gradient Boosting. In addition, these algorithms offer advantages in interpretability, computational efficiency, and their extensive application in material science and energy-related prediction tasks, rendering them suitable candidates for the present study [15–20].

The implementation of the models was conducted using Python, with the scikit-learn and LightGBM libraries utilized. Each model employed “Voltage” and “Light” as input features to predict the “Current” value. Model hyperparameters were maintained at their default settings, except for *n_estimators*, which was fixed at 100 for all models.

2.3.1 Random forest

The Random Forest (RF) algorithm represents a powerful ensemble learning approach based on Bagging (Bootstrap Aggregation) that constructs a considerable number of decorrelated decision trees [21]. During the training phase, two sources of randomness

are introduced to ensure low correlation among trees: (1) each tree is developed using a randomly selected bootstrap sample of the training dataset (Bagging). (2) At each node split, only a randomly selected subset of features is considered (feature randomness). Predictions are derived by averaging the outputs of all the constructed trees. Specifically, the final prediction is determined by calculating the mean output across all individual trees. Figure 3 presents a schematic representation of the Random Forest Regression algorithm. This random selection of features and data serves primarily to reduce model variance, while the aggregation of multiple, independent trees significantly enhances the model’s robustness and generalization capabilities, thereby minimizing the risk of overfitting [22–24]. In this study, Random Forest was utilized as a reference model to predict current values measured under varying illumination intensities, offering a high degree of accuracy and inherent resilience against noise.

2.3.2 Gradient boosting

Gradient Boosting (GB) is an ensemble model characterized by the sequential training of weak learners, typically shallow decision trees. Unlike parallel ensemble methods (e.g., Random Forest), GB aims

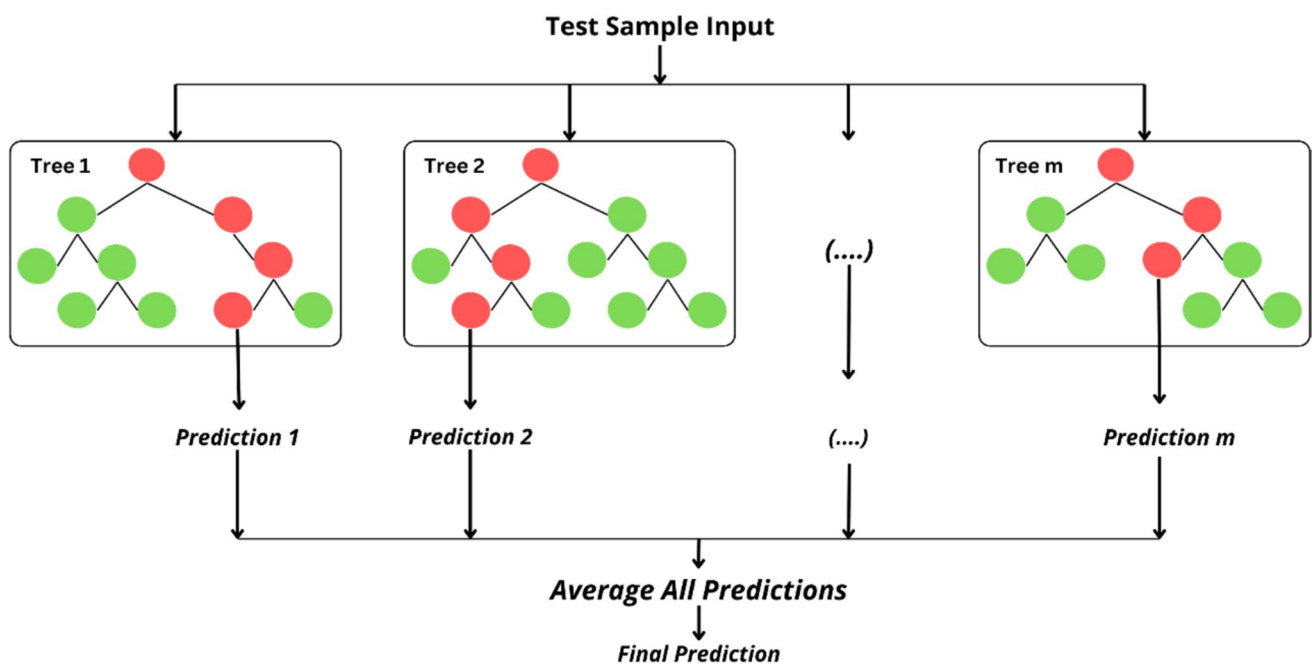


Fig. 3 Random forest regression [25, 26]

primarily to minimize the model's bias. Each newly introduced tree is strategically designed to rectify the errors (residuals) of the preceding cumulative model. This is achieved using a methodology that employs the technique of gradient descent to minimize the loss function. During each iteration, the model prioritizes minimizing these residual errors from prior models, thereby achieving incremental performance enhancements. Figure 4 presents a schematic representation of the Gradient Boosting Regression algorithm. While this sequential learning methodology generally yields high accuracy, it inherently entails greater computational costs due to its lack of parallelizability, and hyperparameter tuning proves to be more sensitive in comparison to parallel ensemble methods [27–29]. This study evaluates the performance of Gradient Boosting in its ability to learn general patterns, with a particular focus on predicting intermediate values, providing a foundational comparison for the more modern boosting implementations (LightGBM and HistGradientBoosting) discussed next.

2.3.3 LightGBM

LightGBM (Light gradient boosting machine) is a gradient boosting algorithm developed by Microsoft, specifically designed to manage large datasets and high-dimensional feature spaces efficiently. It represents an optimization over traditional gradient boosting algorithms. The primary innovation is its histogram-based approach, in which continuous feature

values are discretized into a fixed number of bins. This substantially reduces both memory utilization and computation time during the split-finding process. Furthermore, unlike traditional boosting algorithms that employ a less efficient level-wise growth strategy (growing all nodes at the current depth), LightGBM utilizes a leaf-wise growth strategy. This allows for the formation of deeper, more asymmetric trees that focus more intensely on minimizing the largest residual errors, thereby enhancing the model's ability to capture patterns with greater precision [31, 32]. Figure 5 presents a schematic representation of the LightGBM regression algorithm. This study selected LightGBM due to its rapid learning capabilities and robust generalization performance, even when applied to our limited experimental dataset, making it highly efficient for model prototyping and evaluation.

2.3.4 HistGradientBoosting

HistGradientBoosting, a model developed within the scikit-learn library, represents a highly optimized boosting method that employs a histogram-based learning approach akin to LightGBM. This technique involves partitioning continuous data into a predetermined number of intervals (bins), thereby significantly reducing the memory footprint and enhancing the efficiency of the split-finding and learning process, especially over traditional gradient boosting. Figure 6 presents a schematic representation of the HistGradientBoosting regression algorithm. A significant

Fig. 4 Gradient Boosting schema [30]

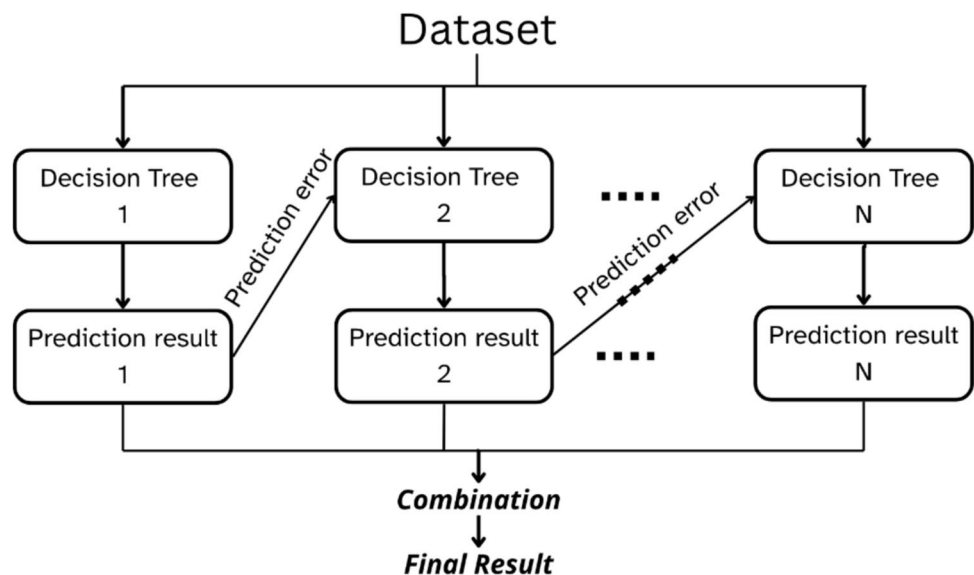


Fig. 5 LightGBM schema [33]

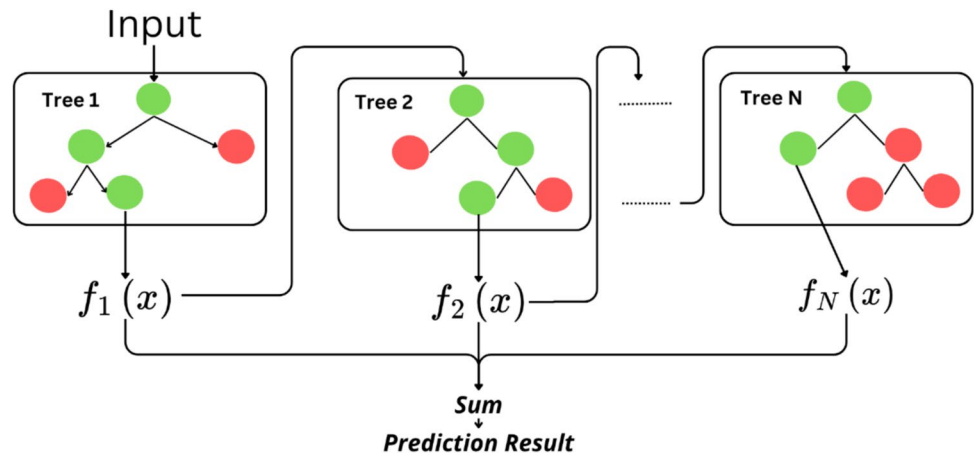
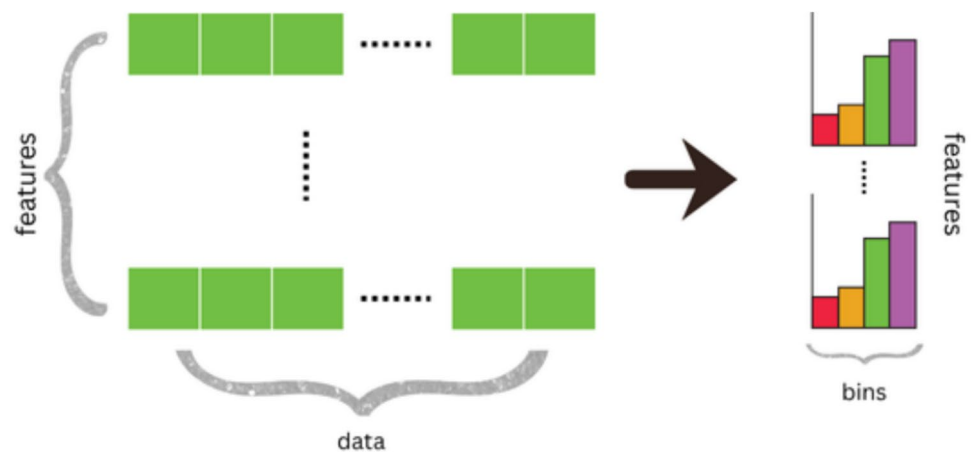


Fig. 6 Histogram-based feature selection for decision trees [38]



advantage of this model is its inherent ability to handle missing values automatically during the training process, treating them as a separate category, and its native support for categorical data encoding. In this study, HistGradientBoosting was selected for its demonstrated potential to optimize both processing speed and accuracy within the established scikit-learn environment, particularly when analyzing a constrained experimental dataset, while offering built-in robustness features [34–37].

2.3.5 ExtraTrees (extremely randomized trees)

The ExtraTrees algorithm exhibits a structure similar to that of the Random Forest method, yet it possesses two important distinctions stemming from its focus on maximizing randomness. First, unlike Random Forest, which utilizes bootstrap sampling (Bagging) in the construction of each tree, ExtraTrees employs the entire original dataset for tree development. Second,

rather than selecting the optimal split thresholds based on criteria (like Gini impurity or variance reduction) at each node, ExtraTrees randomly determines these split thresholds. This increased level of randomization significantly accelerates the learning process by eliminating the time spent searching for the optimal split, while also contributing to a further reduction in model variance and mitigating the likelihood of overfitting. Figure 7 presents a schematic representation of the ExtraTrees Regression algorithm. Generally, this methodology leads to more balanced performance, particularly when addressing high-variance datasets, by favoring higher randomization over meticulous optimization. This model was incorporated into the study due to its potential to generalize the relationships among varying illumination levels effectively and its inherent robustness [39, 40].

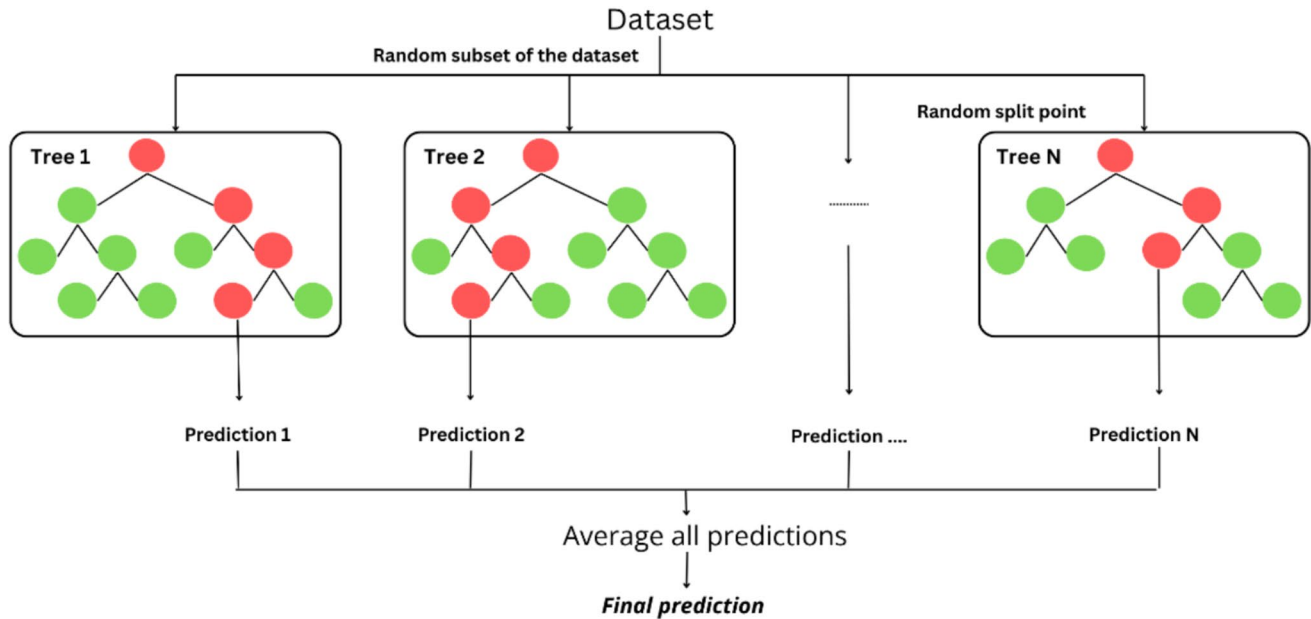


Fig. 7 ExtraTrees [41]

2.4 Model performance evaluation

To assess the predictive performance of the models, four pivotal statistical metrics were employed: R-squared (R^2), mean squared error (MSE), mean absolute error (MAE), and mean absolute percentage error (MAPE). These metrics quantitatively evaluate the discrepancies between predicted and observed values, thereby providing a foundation for comparing model performance.

R^2 signifies the proportion of variance in the dependent variable accounted for by the model, with values approaching 1 indicating greater predictive accuracy [42]. MSE computes the average of the squared prediction errors, imposing a heavier penalty on larger errors and thereby reflecting the overall accuracy of the model. MAE represents the mean of the absolute differences between predicted and actual values, offering straightforward interpretability while being sensitive to the units of measurement. MAPE communicates errors as a percentage, emphasizing the relative magnitude of prediction errors and serving as a handy tool for comparing values across various scales.

By conducting a joint analysis of these four metrics, one can comprehensively evaluate the overall performance of the models as well as the distribution of errors. This method facilitates an assessment of

both the accuracy and generalizability of each model [43–45].

$$R^2 = 1 - \frac{\sum_i^N (y_i - \hat{y}_i)^2}{\sum_i^N (y_i - \bar{y})^2}, \quad (1)$$

$$\text{MSE} = \frac{1}{N} \sum_i^N (y_i - \hat{y}_i)^2, \quad (2)$$

$$\text{MAE} = \frac{1}{N} \sum_i^N |y_i - \hat{y}_i|, \quad (3)$$

$$\text{MAPE} = \frac{1}{N} \sum_i^N \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100, \quad (4)$$

where N is the number of observations, y_i is the actual value, $\hat{y}_i$ is the predicted value by the model, and $\bar{y}$ is the mean of the actual values.

2.5 Statistical validation and ANOVA-based error analysis

2.5.1 ANOVA (analysis of variance)

Analysis of Variance (ANOVA) is a statistical technique used to compare means across multiple groups

and determine whether observed differences are statistically significant [46]. It decomposes the total variability of the data into components attributable to different sources, such as factors of interest and residual error. A key advantage of ANOVA over multiple t tests is that it controls the Type I error rate when comparing more than two groups. In the context of predictive modeling, ANOVA can be applied to evaluate whether the performance metrics (e.g., MSE, R^2 , or log-transformed errors) differ significantly among models or experimental conditions [47].

Two-way ANOVA extends this approach to evaluate the simultaneous effect of two factors and their interaction on the response variable. In our study, the two factors are machine learning model and light intensity, and the response is the log-transformed squared prediction error. A significant interaction indicates that the effect of one factor depends on the level of the other factor [48].

2.5.2 Tukey HSD (*honestly significant difference*) test

The Tukey HSD test is a post hoc pairwise comparison method used after an ANOVA to identify which specific group means differ from each other while controlling the family-wise error rate [49]. It calculates confidence intervals for all pairwise differences between group means and determines whether these intervals include zero. If the interval does not include zero, the difference is considered statistically significant.

In machine learning model evaluation, Tukey HSD can be used to compare performance metrics across multiple algorithms, providing a rigorous way to identify which models outperform others under the same experimental conditions [50]. By applying Tukey HSD to the ANOVA results, we determined that the ExtraTrees model significantly outperforms other models in terms of prediction error across the evaluated illumination levels.

2.6 Workflow of the experimental and machine learning framework

Figure 8 illustrates the overall workflow adopted in this study, encompassing material preparation, experimental data acquisition, machine learning modeling, and statistical validation. Initially, the Au/n-SiO₂ device with a graphene oxide (GO) and polymer-based interface layer was fabricated, followed by

current–voltage measurements under varying illumination levels. The collected dataset was then structured by considering voltage and light intensity as input features and current as the target variable. Multiple ensemble-based regression models were trained to predict intermediate illumination responses. Model performance was quantitatively assessed using standard error metrics, while statistical robustness and significance were further examined through two-way ANOVA and post hoc Tukey HSD analyses. Finally, the results were visualized and interpreted to evaluate the predictive reliability and applicability of machine learning approaches for optoelectronic material characterization.

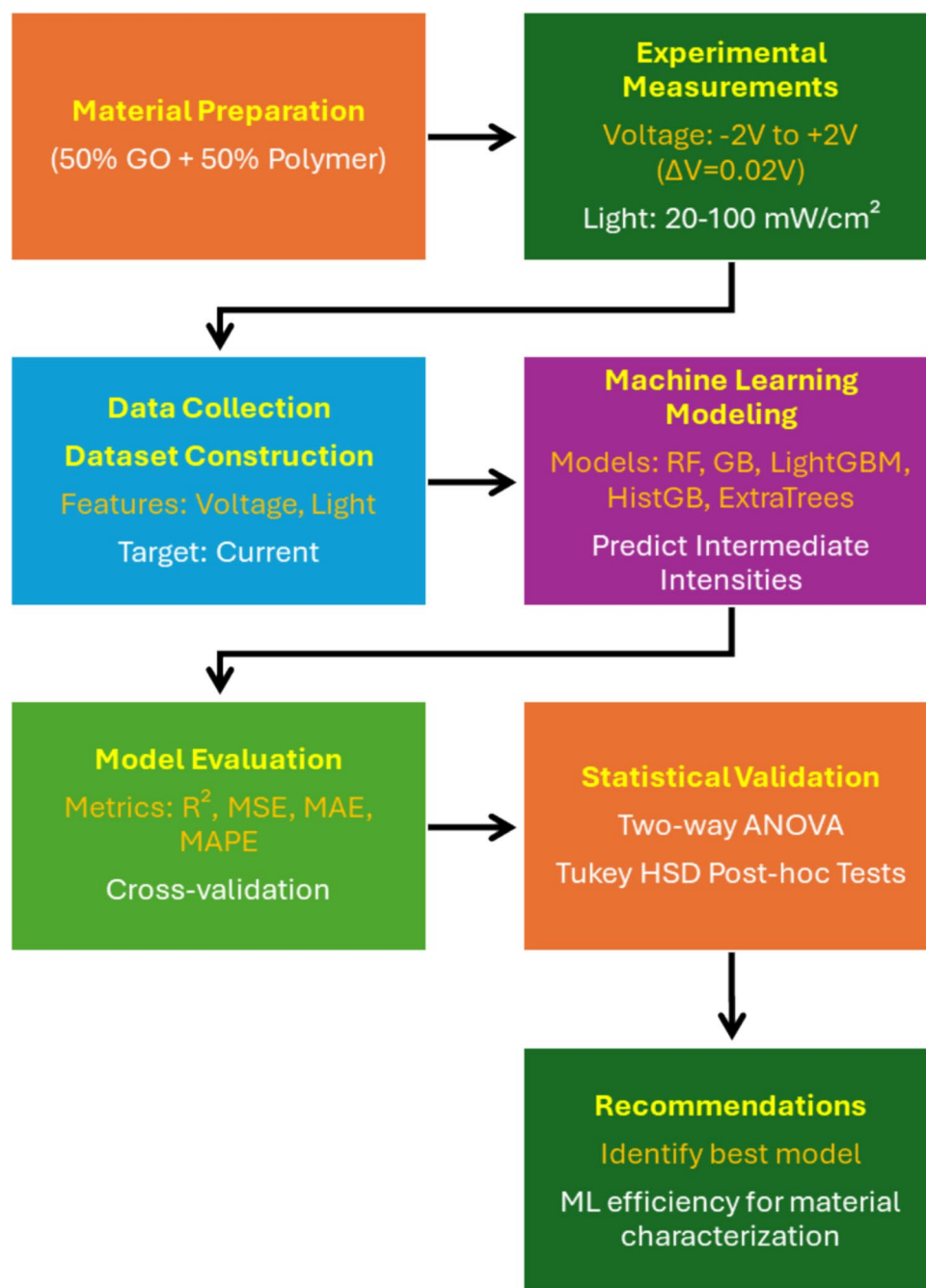
3 Results and discussion

The collected experimental data were utilized to develop and evaluate five machine learning algorithms. The objective of these models was to predict the system response at intermediate illumination levels for which no experimental data were available. The resulting predictions facilitate a comprehensive comparison of the efficacy of different algorithms in the context of polymer-based energy systems.

Model performance was assessed through two complementary approaches.

The first approach entailed evaluating models trained via cross-validation against a distinct test dataset. In this methodology, each model underwent training and testing across all folds, with the average of the performance metrics computed to ascertain overall predictive accuracy. This approach addresses both the model's fit to the training data and its generalization capability on previously unseen test data, thereby mitigating the risk of overfitting and enhancing the reliability of predictions derived from experimental datasets [51, 52].

During the model evaluation process, a fivefold cross-validation was conducted utilizing the StratifiedKFold technique. Although StratifiedKFold is typically employed in classification, its adaptation here ensured that the distribution of the target variable Current across key operating ranges was preserved when forming training and test subsets. This allowed for a more equitable assessment of the predictive performance of the models across varying illumination levels. For each fold, the models were trained, and

Fig. 8 Workflow of the experimental and machine learning framework**Table 1** Performance metrics of regression models

Models	R^2 _Mean	MSE_Mean	MAE_Mean	MAPE_Mean (%)
Random Forest	0.999307	3.79E-12	9.07E-07	11.29508
Gradient Boosting	0.999253	4.14E-12	1.15E-06	13.94759
LightGBM	0.999259	4.1E-12	9.77E-07	13.90074
HistGradientBoosting	0.999322	3.72E-12	9.2E-07	18.68602
ExtraTrees	0.999872	7.12E-13	3.93E-07	6.806718

predictions were generated based on the respective test set.

Table 1 presents the prediction scores of the different regression models based on experimental data. The evaluation was conducted using the average values of four performance metrics: R^2 , MSE, MAE, and MAPE. The inclusion of multiple metrics ensured a balanced evaluation; R^2 measures the goodness of fit, while MSE and MAE quantify error magnitude, and MAPE provides a scale-independent percentage error.

According to Table 1, the ExtraTrees algorithm demonstrated the highest R^2 value (0.999872), signifying that it accounts for nearly all the variance in the dependent variable. This highly favorable result, confirmed by the cross-validation methodology, suggests a strong generalization capability rather than simple overfitting. Additionally, it recorded an MSE of 7.12×10^{-13} , which reflects a strong sensitivity to squared errors. Its MAE and MAPE values are also considerably lower than those of the other models, indicating that both the average magnitude of errors and the relative percentage errors are minimized.

Although the HistGradientBoosting model exhibits commendable performance regarding R^2 , its relatively high MAPE value of 18.68% suggests that it may produce larger errors in some instances, potentially due to the MAPE metric's known sensitivity to prediction errors near zero target values. The Random Forest, LightGBM, and Gradient Boosting models generally present similar performance levels, yielding satisfactory results, particularly with respect to R^2 and MSE metrics. Nevertheless, ExtraTrees appears to provide the most balanced and robust performance across all evaluated metrics. These findings underscore the reliability of the ExtraTrees algorithm as a predictive tool, particularly in contexts where collecting experimental data incurs significant costs.

In the second approach (leave-one illumination-out), a specific illumination level was designated as the test set, while the remaining levels served as the training model. For instance, when 100 mW/cm² was selected as the test set, the other levels (20, 40, 60, and 80 mW/cm²) were utilized for training. This configuration enables a comprehensive evaluation of the model's general performance under an unseen illumination level. While the number of illumination levels is limited, the leave-one illumination-out approach provides a stringent validation scenario by forcing the model to predict entirely unseen intensity conditions. This setup mimics real-world use cases where intermediate

measurements are unavailable. However, increasing the number of illumination levels would further enhance robustness.

The prediction error metrics of the machine learning models at various levels of light intensity are presented in Fig. 9. This evaluation enables a comparative analysis of the models' ability to generalize unknown illumination levels across various algorithms.

Table 2 provides a comprehensive overview of the R^2 , MSE, MAE, and MAPE (%) values for Random Forest, Gradient Boosting, LightGBM, HistGradientBoosting, and ExtraTrees. These performance metrics are assessed across various illumination levels (100, 80, 60, 40, and 20 mW/cm²), enabling the identification of the illumination levels at which each model demonstrates optimal performance. It is noteworthy that this evaluation did not incorporate cross-validation; instead, one illumination level was utilized as the test set, while the remaining levels served as the training set.

According to the findings presented in Table 2, the ExtraTrees algorithm demonstrates exceptional performance, evidenced by high R^2 values (e.g., 0.9993 at 60 mW/cm²) and low error rates across various illumination levels. Notably, the algorithm's low MAPE values indicate that it yields predictions that are most closely aligned with the experimental data. In contrast, Gradient Boosting models exhibited instability in certain instances, with MAPE values exceeding 300% at 40 mW/cm², indicating a considerable level of variance. This comparative analysis provides a valuable reference for identifying the most reliable predictive models in systems that demand precise forecasting, particularly within energy storage applications.

Under conditions of maximum illumination (100 mW/cm²), all models achieved a remarkably high prediction accuracy. The Random Forest model recorded the highest R^2 value (0.9992) and also exhibited the lowest Mean Squared Error (MSE). However, with respect to MAPE, LightGBM, and HistGradientBoosting recorded lower percentage errors (9.32 and 9.25%, respectively), indicating that while overall accuracy remains high across the models, different metrics emphasize distinct advantages inherent to each model. The high quality and discriminative nature of the data under intense illumination likely contributed to the consistently elevated accuracy observed across all models.

Upon reducing the illumination level to 80 mW/cm², a discernible decline in model performance was noted.

Fig. 9 Prediction error metrics of machine learning models for unseen illumination levels

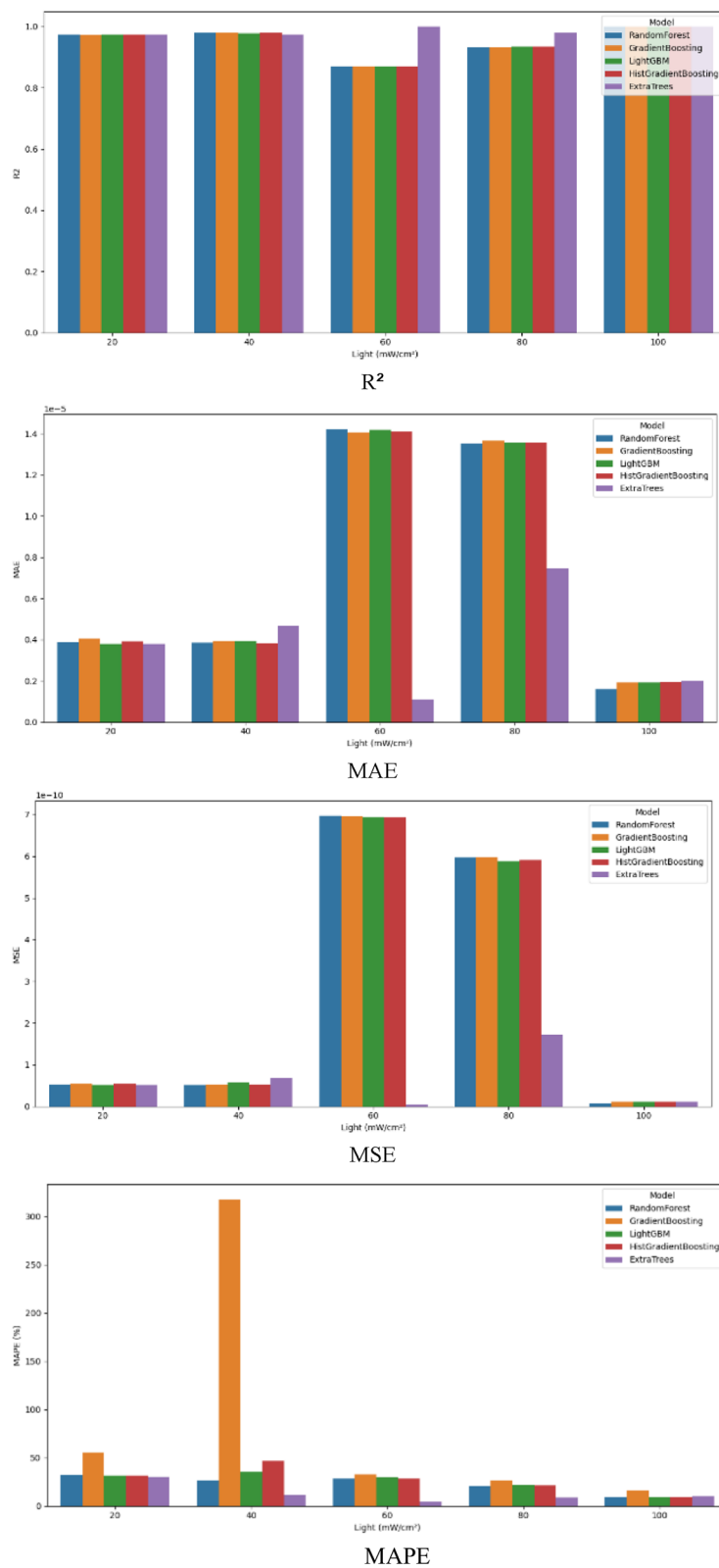


Table 2 Error rates in models

Models	Light (Test)	R^2	MSE	MAE	MAPE (%)
ExtraTrees	100	0.998773	1.09E-11	1.99E-06	9.984776
Gradient Boosting	100	0.998752	1.11E-11	1.92E-06	16.16175
HistGradientBoosting	100	0.998815	1.05E-11	1.94E-06	9.245373
LightGBM	100	0.998752	1.11E-11	1.92E-06	9.320473
Random Forest	100	0.999218	6.94E-12	1.6E-06	9.185236
ExtraTrees	80	0.980465	1.72E-10	7.46E-06	8.721158
Gradient Boosting	80	0.932081	5.98E-10	1.37E-05	26.33528
HistGradientBoosting	80	0.932919	5.91E-10	1.36E-05	21.55016
LightGBM	80	0.933244	5.88E-10	1.36E-05	21.7833
Random Forest	80	0.932291	5.96E-10	1.35E-05	20.66542
ExtraTrees	60	0.999332	3.58E-12	1.09E-06	4.055879
Gradient Boosting	60	0.870053	6.96E-10	1.41E-05	32.57927
HistGradientBoosting	60	0.870527	6.93E-10	1.41E-05	28.46311
LightGBM	60	0.870277	6.94E-10	1.42E-05	29.31916
Random Forest	60	0.869665	6.98E-10	1.42E-05	28.23594
ExtraTrees	40	0.973414	6.87E-11	4.66E-06	11.31749
Gradient Boosting	40	0.979656	5.26E-11	3.92E-06	317.4781
HistGradientBoosting	40	0.979602	5.27E-11	3.83E-06	46.39012
LightGBM	40	0.978018	5.68E-11	3.92E-06	35.04533
Random Forest	40	0.979944	5.18E-11	3.86E-06	26.35738
ExtraTrees	20	0.973996	5.17E-11	3.79E-06	30.38329
Gradient Boosting	20	0.972504	5.46E-11	4.04E-06	55.44272
HistGradientBoosting	20	0.972772	5.41E-11	3.89E-06	31.57342
LightGBM	20	0.974188	5.13E-11	3.8E-06	31.36703
Random Forest	20	0.973385	5.29E-11	3.89E-06	32.27862

Bold font highlights the most favorable metric values, corresponding to the highest R^2 and the lowest error values (MSE, MAE, and MAPE)

At this illumination level, the R^2 values for the Random Forest, Gradient Boosting, and LightGBM models were approximately 0.93, whereas the ExtraTrees model maintained a pronounced advantage with an R^2 of 0.980. In terms of MAPE, ExtraTrees again achieved the lowest error at 8.72%, further confirming its robustness under moderately high-illumination conditions.

At an illumination level of 60 mW/cm², the ExtraTrees model demonstrated a remarkable R^2 value of 0.999, coupled with the lowest MAPE of 4.05%. This performance indicates superior stability and consistency under moderate light conditions. In contrast, other models exhibited more pronounced performance degradation at this level, while ExtraTrees showed greater resilience to variations in illumination.

As the illumination was further reduced to 40 and 20 mW/cm², the overall predictive accuracy of all models declined relative to the baseline established at

100 mW/cm². Although R^2 values generally remained around 0.97, the Gradient Boosting model displayed an exceptionally high MAPE of 317% at an illumination level of 40 mW/cm², indicating substantial prediction errors under these conditions. Conversely, ExtraTrees maintained an R^2 of approximately 0.973, with MAPE values of 11.3% at 40 mW/cm² and 30.3% at 20 mW/cm², thereby demonstrating a strong level of robustness. LightGBM and HistGradientBoosting also produced relatively balanced results. These findings suggest that the overall predictive performance of the models does not simply scale with varying illumination levels but instead relies on their capacity to generalize across different light conditions.

The predictive performance of the five machine learning models at different illumination levels is illustrated in Fig. 10. Figure 10a shows the variation of R^2 with respect to light intensity, indicating that all models maintain high scores at higher illumination levels,

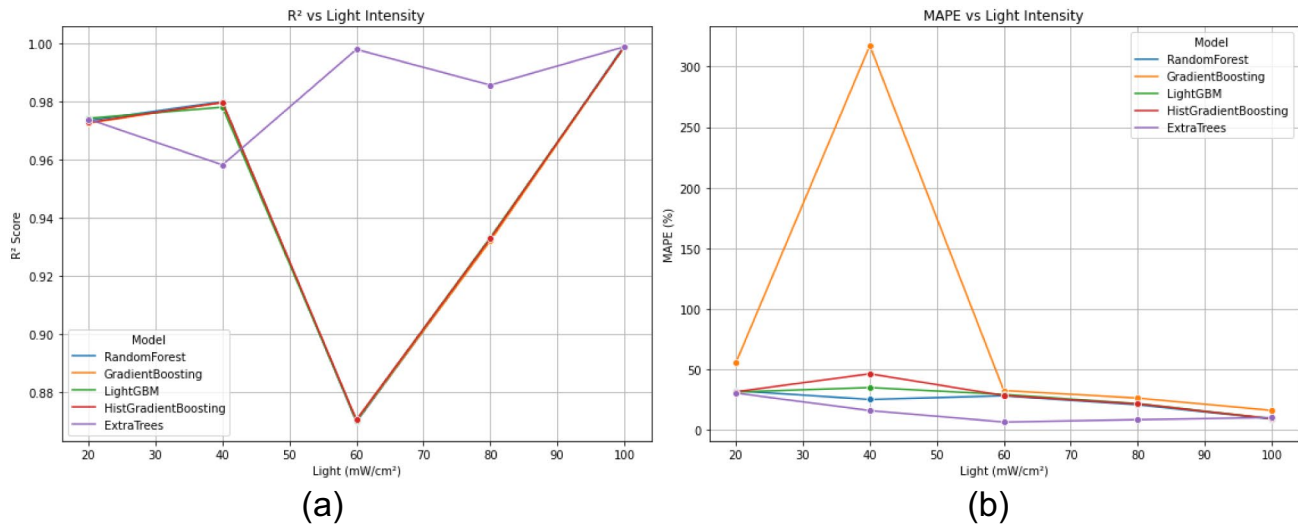


Fig. 10 R^2 and MAPE scores

while performance slightly decreases at lower intensities. Figure 10b presents the MAPE values, which highlight the relative prediction errors across models. The ExtraTrees and Random Forest models consistently achieve the lowest MAPE, demonstrating their robustness in predicting current values under varying light conditions.

In the context of varying illumination levels, the ExtraTrees model demonstrates notable stability and low error rates, distinguishing itself from other models. It consistently performs well in both high- and low-illumination conditions, particularly excelling in key error metrics such as MAE and MAPE. ExtraTrees introduces more randomization, which might have helped in reducing variance over this specific sparse experimental dataset. While both Random Forest and LightGBM exhibit strong performance in high-illumination environments, they experience a decline in effectiveness as illumination levels diminish. Gradient Boosting, on the other hand, shows the most significant reduction in predictability when operating under low-light conditions. Consequently, the ExtraTrees model is recommended for use in environments characterized by variable or low illumination, whereas all models generally yield reliable results in conditions of high illumination.

In Fig. 11, the assessment of model stability was conducted by calculating the standard deviation for each error metric. The analysis revealed that Gradient Boosting exhibited the highest variability in the MAPE, while Random Forest demonstrated the

lowest variability. Conversely, for R^2 , MSE, and MAE, ExtraTrees presented the lowest variability, in contrast to Random Forest, which exhibited the highest.

Figure 12 presents the comparison of actual and predicted distributions for each regression model, illustrating the extent to which the models reflect the underlying statistical behavior of the experimental data across different illumination levels.

Figure 13 displays experimental measurements alongside the predicted values for intermediate illumination levels. The predictions corresponding to these intermediate points align closely with the experimental data, indicating a level of consistency and reliability in the model outputs. As illustrated in Fig. 13f, the experimental measurement graph demonstrates that the current values of the materials under examination increase with elevated illumination levels for both negative and positive voltage measurements. This phenomenon can be attributed to the enhanced generation of electron-hole pairs resulting from increased light exposure [53, 54]. Previous studies [4, 5] have calculated electrical parameters such as n , Φ_{bo} , R_s , I_0 , and RR for the Au/(GO:P3C4MT)/SiO₂ structure using the I - V graphs mentioned above. These calculations employed Thermionic Emission (TE), Cheung, and Norde methods across varying illumination levels, with the results subjected to comparative analysis.

Although tree-based models do not directly yield physical parameters such as the ideality factor (n) or series resistance (R_s), they provide a high-fidelity representation of the Current-Voltage-Light (I - V - L)

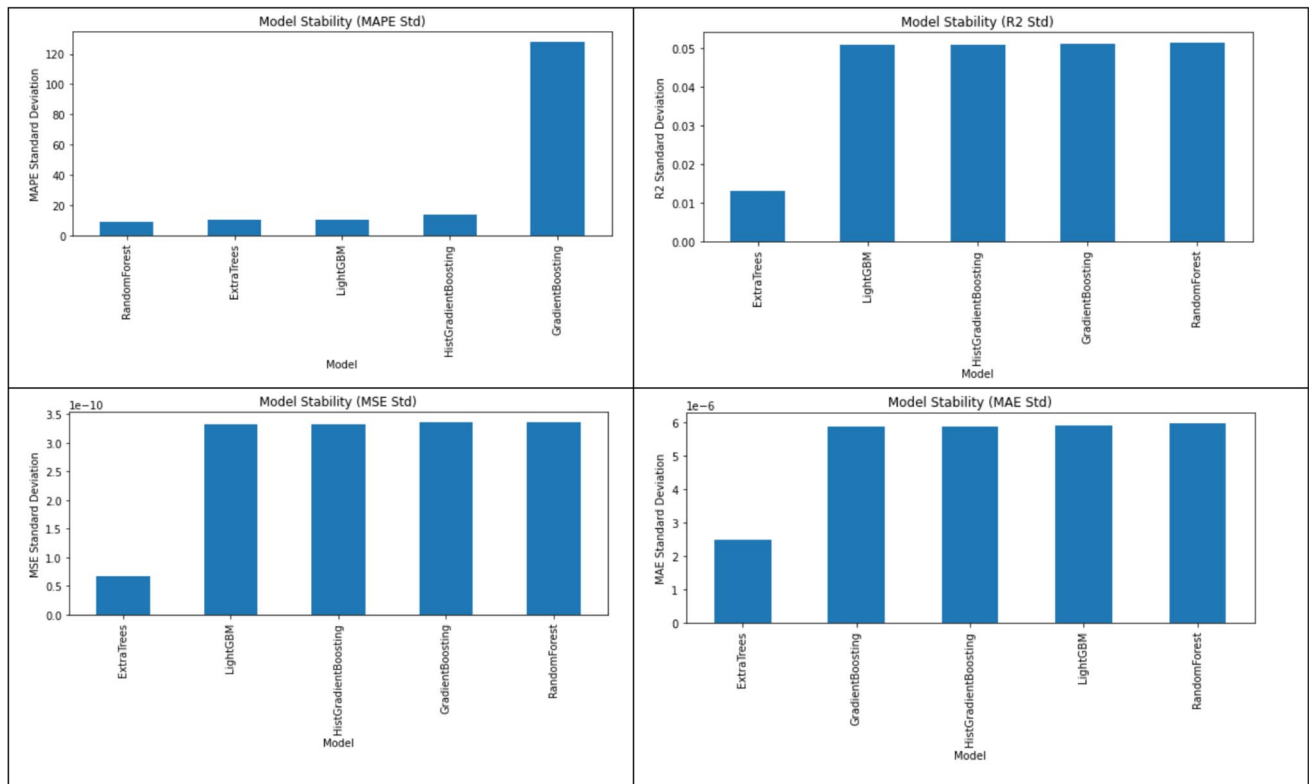


Fig. 11 Model stability based on error metrics

relationship by learning the underlying complex dependencies between voltage, illumination, and current. This allows characteristic device behaviors, including turn-on voltage shifts and saturation effects, to emerge naturally from the data without explicit physical parameterization.

The strong predictive performance, particularly of the ExtraTrees model, indicates that the reconstructed I – V characteristics are consistent with key optoelectronic properties. For instance, the high accuracy in the low-forward bias region ($V < V_{on}$) reflects behavior associated with the ideality factor and barrier height, while steeper slopes at higher voltages under increased illumination suggest a reduction in effective series resistance due to enhanced carrier generation.

By generating smooth and noise-reduced I – V curves, the model enables indirect analysis of critical device parameters and supports subsequent physics-based extraction methods. This approach reduces experimental effort while preserving physical interpretability, offering a complementary alternative to conventional modeling techniques [55, 56].

Based on the computed electrical parameters, the material under study exhibits potential applicability in

electronic circuits, rectifying contacts, and integrated circuit applications. The observed increase in current with heightened illumination confirms that the Au/(GO:P3C4MT)/SiO₂/n-Si material is photosensitive. This finding suggests that the Au/(GO:P3C4MT)/SiO₂/n-Si structure possesses a commendable rectification ratio and linearity, indicating its promising potential for future photovoltaic and photodiode applications.

Predictions derived from the Gradient Boosting machine learning method indicate an increase in current values with higher illumination at negative voltage levels. However, the predicted values for consecutive points appear either identical or demonstrate minimal variation, as evidenced by the plotted graph. This behavior is unexpected. For positive voltage predictions, the calculated current values are closely aligned, causing the graph to exhibit overlapping characteristics within the initial voltage range (0–0.5 V). Additionally, deviations from experimental data are noted at other voltage levels, suggesting that the method does not accurately predict positive voltages.

Conversely, predictions generated using the ExtraTrees method reveal a proportional increase in

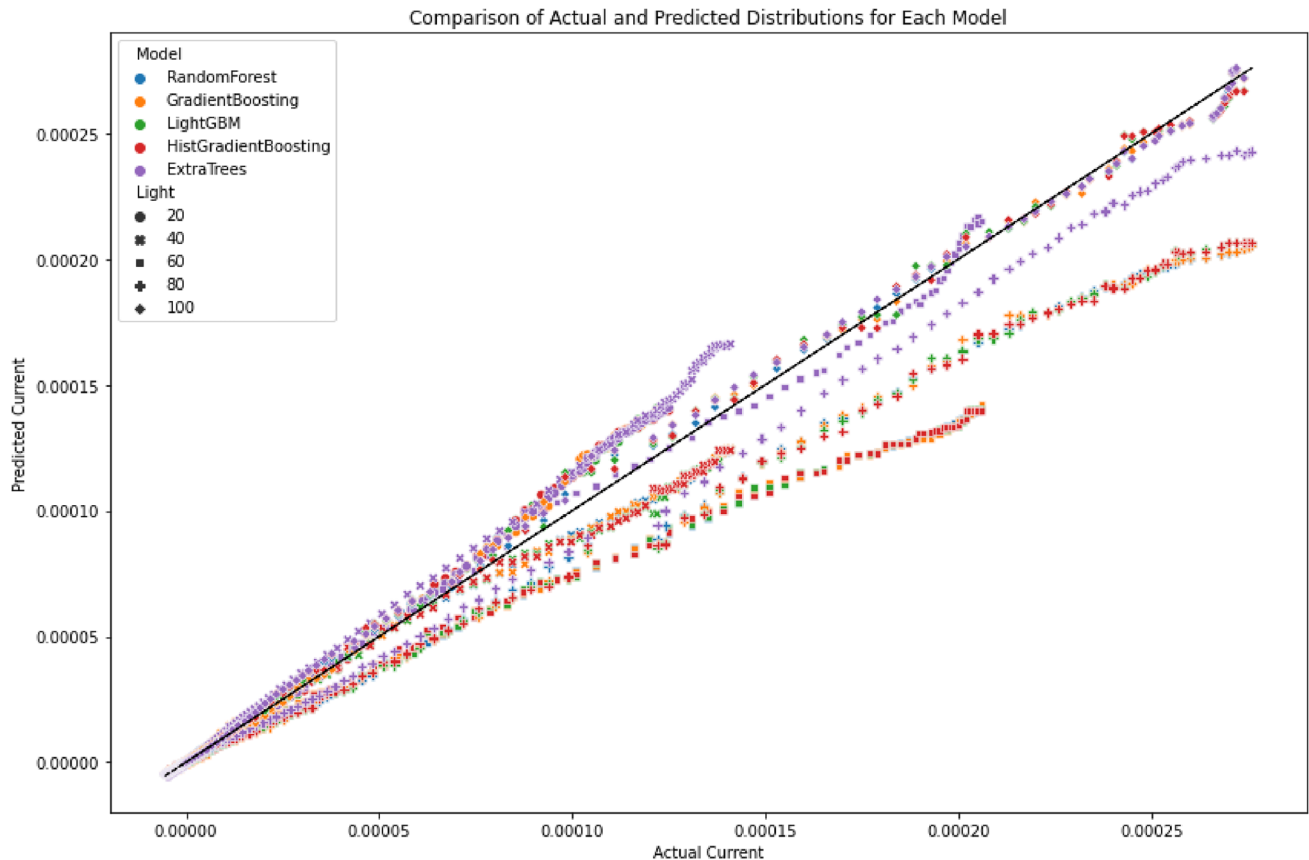


Fig. 12 Comparison of actual and predicted distributions

current values with rising illumination under negative voltages (Fig. 13b). This result is consistent with expectations and aligns well with findings reported in the literature [57, 58].

The predictions derived from the ExtraTrees method demonstrate a close alignment with experimental results for positive voltage values. This observation signifies that this method is the most effective among those evaluated in the study. Given the accurate correspondence between predicted values and experimental data, the ExtraTrees method presents significant promise as a tool for assessing the optoelectronic properties, such as photodiode and photovoltaic characteristics, of the Au/(GO:P₃C₄MT)/SiO₂/n-Si structure, thus reducing the need for extensive laboratory procedures.

In a similar context, predictions generated by the Random Forest method indicate a proportional increase in current in response to rising illumination levels at negative voltages, thereby affirming the reliability of this methodology. Furthermore, for positive voltage values, Random Forest predictions

closely resemble the experimental data; however, the accuracy is slightly inferior when compared to that of the ExtraTrees method. Therefore, Random Forest is regarded as the second most effective regression method among those utilized in this investigation.

Predictions obtained through the LightGBM and HistGradientBoosting methods also indicate that current values rise proportionally with increasing illumination at negative voltages [59]. However, the predictions for intermediate values closely approximate the experimental measurements, resulting in slight discontinuities within the plotted curves. This phenomenon suggests that predicted values from these methods do not exhibit linear progression relative to their predecessors. A similar pattern is noted for positive voltage values, indicating that while these methods provide reliable predictions, they do not attain the optimality of the Random Forest and ExtraTrees methodologies. Nonetheless, through the application of each machine learning method, it is possible to derive detailed optoelectronic parameters—such as n , Φ_{bo} , R_s , I_0 , and RR for the Au/(GO:P₃C₄MT)/SiO₂

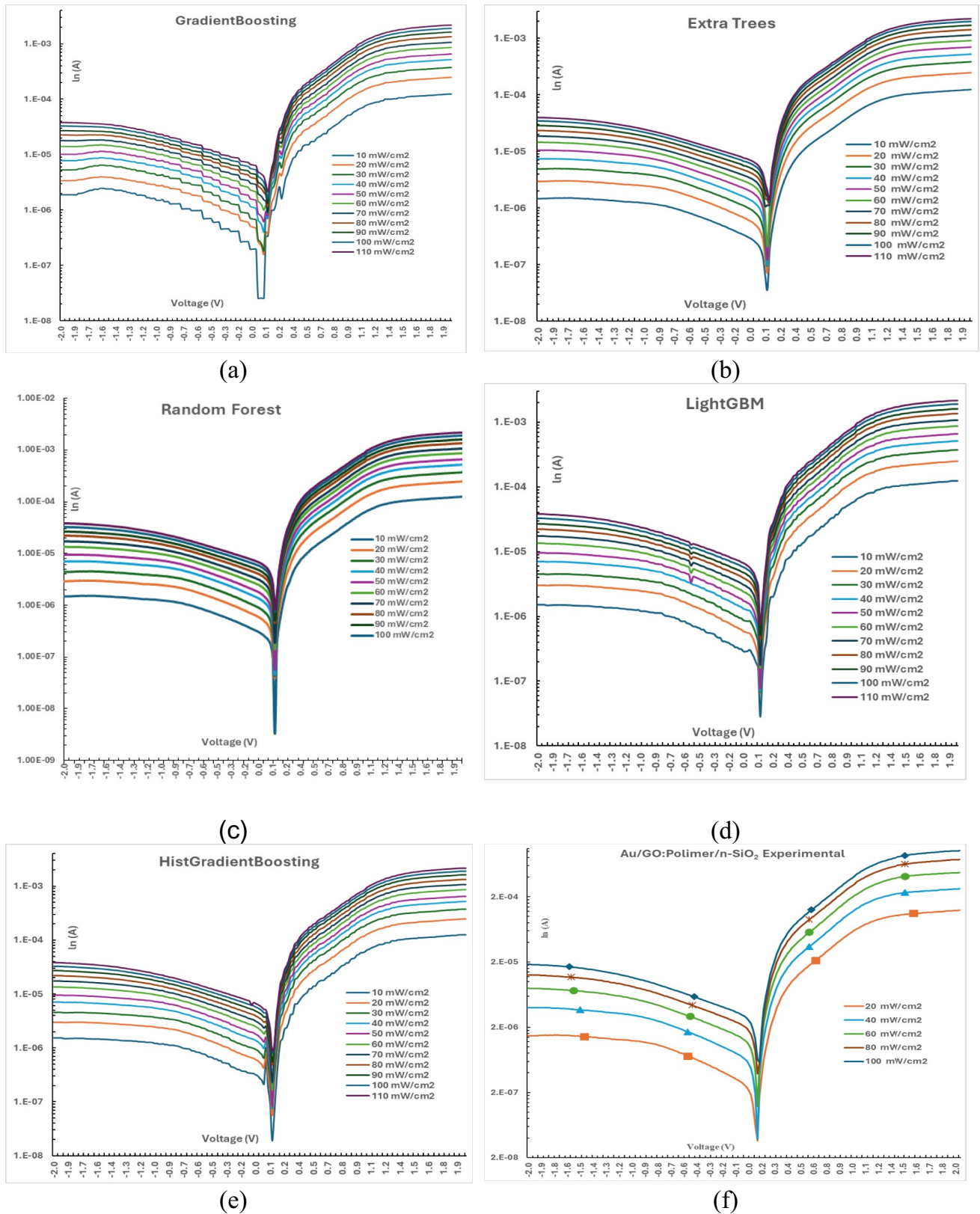


Fig. 13 Machine learning predictions and experimental data graphs. **a** Gradient Boosting, **b** ExtraTrees, **c** Random Forest, **d** LightGBM, **e** HistGradientBoosting, and **f** Experimental data

structure—from the predicted electrical and optical data. These data may serve as a valuable foundation for the design of efficient photovoltaic and photodiode devices, thereby underscoring the practical implications of this research in the field of optoelectronics.

The results presented above were obtained without optimizing the hyperparameters. To more accurately assess the models' effectiveness, selecting appropriate hyperparameters is crucial for obtaining valuable insights. Each model was assessed using a variety of hyperparameter ranges, with optimal parameter combinations systematically explored through GridSearchCV utilizing fivefold cross-validation (fivefold CV). Specific parameter ranges were defined for each machine learning model. For both Random Forest and ExtraTrees, the number of trees (*n_estimators*) was tested at 100 and 200, while *max_depth* was set to none (unrestricted), 10, and 20. In the case of Gradient Boosting and LightGBM, to balance the model's learning capacity with generalization ability, *n_estimators* was evaluated at 100 and 200, and the learning rate was set at 0.05 and 0.1.

Furthermore, to mitigate overfitting, *max_depth* was set to 3 and 5 for Gradient Boosting, and to -1 (unlimited depth) and 10 for LightGBM. For HistGradientBoosting, the *max_iter* (total number of iterations) was tested at 100 and 200, with learning rates of 0.05 and 0.1, and *max_depth* set to none (unrestricted) and 10. These parameter ranges were carefully selected to ensure that the models maintain sufficient flexibility while remaining computationally feasible, considering the size of the experimental dataset. This configuration enabled an evaluation of each model under conditions that optimized its performance, with the results detailed in Table 3.

An analysis of Table 3 indicates that all models exhibited very high accuracy in relation to the experimental data. In particular, the ExtraTrees algorithm

achieved an R^2 score of 1.000, reflecting near-perfect agreement and exhibiting the lowest error values (MSE: 3.91×10^{-18} , MAE: 7.14×10^{-10} , MAPE: 0.0685%). This performance underscores its exceptional ability to accurately model the light-current relationship and produce highly reliable predictions.

The Gradient Boosting model also demonstrated high levels of accuracy with optimized hyperparameters (*learning_rate*: 0.1, *n_estimators*: 200, *max_depth*: 5; R^2 : 0.999994). However, its MAPE value of 4.26% was higher than that of the ExtraTrees algorithm. Furthermore, the LightGBM, HistGradientBoosting, and Random Forest models yielded commendable results, with R^2 scores approximating 0.9998 and MAPE values ranging from 1.9 to 2.6%.

Overall, hyperparameter optimization significantly enhanced the predictive performance of the models, with a particular emphasis on the ExtraTrees algorithm due to its low error and high overall accuracy. This optimization also markedly improved the prediction accuracy of the Gradient Boosting model.

To further validate the necessity of ensemble learning, the performance of the ExtraTrees model was compared against traditional polynomial regression. Polynomial regression (degree = 3) was implemented as a baseline parametric approach and evaluated under the same leave-one-illumination-out validation strategy. The results indicate that Polynomial Regression achieves moderate predictive performance, with R^2 values ranging from 0.568 to 0.964 depending on the illumination level. While acceptable performance is observed at intermediate intensities (e.g., $R^2 \approx 0.96$ at 60–80 mW/cm²), a significant degradation is evident at lower and higher illumination levels ($R^2 \approx 0.57$ at 20 and 0.88 at 100 mW/cm²). This behavior suggests that polynomial models struggle to generalize across the full nonlinear response range, particularly under low-signal conditions where noise and nonlinearity

Table 3 Hyperparameter optimization results of machine learning models

Models	Best parameters	R^2	MSE	MAE	MAPE (%)
Random Forest	'max_depth': None, 'n_estimators': 100	0.999919	4.57E-13	3.11E-07	1.916218
Gradient Boosting	'learning_rate': 0.1, 'max_depth': 5, 'n_estimators': 200	0.999994	3.50E-14	1.13E-07	4.268977
LightGBM	'learning_rate': 0.1, 'max_depth': -1, 'n_estimators': 200	0.999861	7.92E-13	3.80E-07	2.676973
HistGradientBoosting	'learning_rate': 0.1, 'max_depth': None, 'max_iter': 200	0.999844	8.84E-13	4.28E-07	2.442482
ExtraTrees	'max_depth': None, 'n_estimators': 200	1	3.91E-18	7.14E-10	0.068516

Bold font highlights the most favorable metric values, corresponding to the highest R^2 and the lowest error values (MSE, MAE, and MAPE)

are more pronounced. Unlike polynomial regression, which assumes a fixed global functional form, tree-based ensemble methods adaptively capture localized nonlinearities in the current–voltage–light relationship without requiring prior assumptions.

A promising alternative to purely data-driven models is physics-informed machine learning (PIML), which integrates governing physical equations such as the Poisson equation or the Shockley diode model into the learning process. These approaches enforce physical consistency and may reduce data requirements [60, 61]. Conventional analytical models, particularly the Shockley diode equation, have long been employed to describe the current–voltage behavior of semiconductor devices due to their clear physical interpretability. However, their applicability becomes limited when dealing with hybrid structures that exhibit complex and non-ideal transport mechanisms. In such systems, phenomena such as recombination losses, interface states, and illumination-dependent barrier height variations introduce deviations from ideal diode behavior. These effects are difficult to fully capture within a single, simplified analytical framework, leading to reduced predictive accuracy. In contrast, the proposed ExtraTrees and other machine learning frameworks provide a flexible and data-driven alternative capable of modeling these complex, nonlinear interactions without relying on restrictive physical assumptions. This makes it particularly suitable for hybrid semiconductor systems where conventional models fail to adequately represent the underlying physics.

Tukey HSD test was conducted to evaluate pairwise differences in prediction performance among the machine learning models, while controlling the family-wise error rate (FWER = 0.05). FWER represents the probability of making at least one Type I error across all pairwise comparisons, ensuring that the overall risk of false positives remains below 5%. Table 4 presents the results, where the columns are defined as follows: Group 1/Group 2 indicate the two models being compared. Mean Difference (meandiff) shows the average difference in the performance metric between the groups, with positive values indicating higher error for Group 1. Adjusted p value (p -adj) represents the significance level corrected for multiple comparisons, with values below 0.05 indicating statistically significant differences. Lower CI/Upper CI provide the 95% confidence interval for the mean difference; and Reject H_0 denotes whether the null hypothesis of no difference is rejected.

The results indicate that the ExtraTrees model achieves significantly lower prediction errors compared to all other models, demonstrating its superior accuracy and consistency. In contrast, the performance differences among Gradient Boosting, HistGradientBoosting, LightGBM, and Random Forest were not statistically significant ($p > 0.98$), suggesting comparable performance among these models. Consequently, under conditions of limited data and resources, the ExtraTrees model is recommended to optimize prediction accuracy and reduce experimental verification efforts, while the other models remain viable alternatives with reasonable performance.

The analysis in Table 5 reveals that the highest illumination level (100 mW/cm²) differs significantly

Table 4 Model performance comparisons with Tukey HSD (FWER = 0.05)

Model group 1	Model group 2	Mean difference (meandiff)	Adjusted p value (p -adj)	Lower CI	Upper CI	Reject H_0
ExtraTrees	Gradient Boosting	0.4165	0.0019	0.1104	0.7226	True
ExtraTrees	HistGradientBoosting	0.3727	0.0080	0.0666	0.6788	True
ExtraTrees	LightGBM	0.3739	0.0077	0.0678	0.6800	True
ExtraTrees	Random Forest	0.3627	0.0108	0.0565	0.6688	True
Gradient Boosting	HistGradientBoosting	−0.0438	0.9951	−0.3499	0.2623	False
Gradient Boosting	LightGBM	−0.0426	0.9956	−0.3487	0.2635	False
Gradient Boosting	Random Forest	−0.0538	0.9892	−0.3599	0.2523	False
HistGradientBoosting	LightGBM	0.0012	1.0000	−0.3049	0.3073	False
HistGradientBoosting	Random Forest	−0.0100	1.0000	−0.3161	0.2961	False
LightGBM	Random Forest	−0.0112	1.0000	−0.3173	0.2949	False

Table 5 Tukey HSD pairwise comparison of light intensity levels (FWER = 0.05)

Light group 1	Light group 2	Mean difference (mean-diff)	Adjusted p value (p -adj)	Lower CI	Upper CI	Reject H_0
100	20	0.4105	0.0016	0.1121	0.7089	True
100	40	0.4884	0.0001	0.1900	0.7868	True
100	60	1.3235	0.0000	1.0251	1.6219	True
100	80	1.5362	0.0000	1.2378	1.8346	True
20	40	0.0779	0.9537	-0.2205	0.3763	False
20	60	0.9130	0.0000	0.6146	1.2114	True
20	80	1.1257	0.0000	0.8274	1.4241	True
40	60	0.8351	0.0000	0.5367	1.1335	True
40	80	1.0478	0.0000	0.7494	1.3462	True
60	80	0.2127	0.2934	-0.0857	0.5111	False

Table 6 Two-way ANOVA results for model and light effects

Sources	Sum of squares (sum_sq)	df	F value	p value
Model (C(Model))	118.6897	4	5.0133	4.956×10^{-4}
Light (C(Light))	1701.6816	4	71.8762	2.720×10^{-59}
Model $\times$ Light	448.1795	16	4.7326	1.176×10^{-9}
Residual	29,593.9804	5000	—	—

from most of the lower intensities (20, 40, 60, 80 mW/cm²), with p -adj < 0.05, indicating a substantial effect of light intensity on prediction performance. Similarly, mid-range intensities (20, 40, 60, 80 mW/cm²) exhibit significant differences among themselves, except for some pairs such as 20 vs. 40 and 60 vs. 80, where p -adj > 0.05, suggesting comparable predictive errors at these levels. These results confirm that light intensity is a critical factor influencing the model performance, highlighting the importance of evaluating model accuracy across the full range of experimental illumination conditions.

The results of the two-way ANOVA assessing the effects of machine learning model and light intensity on prediction errors are summarized in Table 6. The table columns are defined as follows: Source indicates the factor or interaction being tested (Model, Light, Model $\times$ Light, or Residual). *Sum of Squares* (sum_sq) represents the total variation attributable to each source; *df* denotes the degrees of freedom associated with the source; *F* value is the test statistic comparing the variance explained by the source to the residual variance; and *p* value indicates the statistical significance of the effect.

The analysis demonstrates significant main effects of model type ($F(4,5000) = 5.01$, $p < 0.001$) and light intensity ($F(4,5000) = 71.88$, $p < 0.001$) on prediction errors, indicating that both factors independently influence model performance. Furthermore, the Model $\times$ Light interaction is also significant ($F(16,5000) = 4.73$, $p < 0.001$), suggesting that the effect of model selection on prediction accuracy depends on the illumination level. In other words, certain models perform relatively better or worse at specific light intensities. These results are consistent with the Tukey HSD post hoc comparisons, which identified significant pairwise differences among models and light levels, collectively providing a robust statistical validation of model performance across experimental conditions.

The comparative analysis of various machine learning algorithms demonstrates that ExtraTrees consistently outperforms its counterparts in terms of stability, accuracy, and error minimization across varying light intensities. While Gradient Boosting, LightGBM, HistGradientBoosting, and Random Forest also deliver reliable predictions, their performance exhibits greater sensitivity to fluctuations in illumination levels. The incorporation of hyperparameter optimization further elevates model performance, particularly for Gradient Boosting, thereby confirming that meticulous tuning can substantially enhance predictive capabilities. Overall, the findings presented highlight the significant potential of machine learning methodologies in precisely estimating optoelectronic properties derived from constrained experimental datasets. This capability not only reduces the burden on laboratory resources but also lays the groundwork

for a comprehensive framework for designing and analyzing optoelectronic devices based on graphene oxide–polymer composites.

4 Conclusion

In this study, ensemble-based machine learning models were successfully applied to predict the optoelectronic behavior of GO:P₃C₄MT hybrid photovoltaic structures under varying illumination conditions. Using experimentally obtained *I*–*V* data, the models demonstrated strong predictive capability not only within the trained range but also at intermediate and extrapolated light intensities.

Among the evaluated methods, the ExtraTrees algorithm exhibited the most consistent and reliable performance across all illumination levels, particularly under low-light conditions where other models showed notable degradation. Statistical analyses confirmed that these performance differences were significant, highlighting the robustness of the proposed approach.

The findings demonstrate that data-driven models can effectively reduce the need for extensive experimental measurements while preserving high predictive accuracy. Although the current implementation is limited to a fixed material ratio, the methodology is inherently extensible to multi-parameter systems. Future work will focus on incorporating additional features such as material composition and hysteresis effects to further enhance model generalizability.

Overall, this study establishes a reliable and scalable framework for machine learning-assisted characterization of optoelectronic materials, offering a practical complement to conventional experimental and analytical approaches.

Author contributions

Ömer Arslan and İhsan Tuğal contributed primarily to the machine learning aspects of the study, including data preprocessing, model development, training, validation, and statistical analysis of the results. Halil Seymen and Şükrü Karataş were mainly responsible for the experimental work, including device

fabrication, electrical measurements, and data acquisition. All authors participated in the interpretation of the results, manuscript writing, and critical revision of the paper. All authors have read and approved the final version of the manuscript.

Funding

Open access funding provided by the Scientific and Technological Research Council of Türkiye (TÜBİTAK). The authors declare that no funds, grants or other support were received during preparation of this manuscript.

Data availability

Data will be made available on request.

Declarations

Conflict of interest The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. R. Özmenteş, Characterization of CuO Thin Film and CuO/n-Si Heterojunction Solar Cells. Mater. Sci.-Pol. (2017). <https://doi.org/10.2478/msp-2018-0092>

2. A. Bingol, O. Arslan, Effective utilization of solar energy in sustainable agricultural irrigation a provincial example from Turkiye. *Therm. Sci.* **27**, 3373–3384 (2023). <https://doi.org/10.2298/TSCI2304373B>
3. H. Özerli, A. Bekereci, A. Türit, Ş Karataş, Electrical and photovoltaic properties of Ag/p-Si structure with GO doped NiO interlayer in dark and under light illumination. *J. Alloys Compd.* **718**, 75–84 (2017). <https://doi.org/10.1016/j.jallcom.2017.05.121>
4. H. Seymen, Ş Karataş, Investigation of electrical, photodiode and photovoltaic properties of Au/SiO₂/n-Si structures with GO and P₃C₄MT interface. *Mater. Chem. Phys.* **310**, 128449 (2023). <https://doi.org/10.1016/j.matchemphys.2023.128449>
5. H. Seymen, N. Berk, İ Orak, Ş Karataş, Effect of illumination intensity on the electrical characteristics of Au/SiO₂/n-type Si structures with GO and P₃C₄MT interface layer. *J. Mater. Sci. Mater. Electron.* **33**, 19656–19666 (2022). <https://doi.org/10.1007/s10854-022-08801-w>
6. A.M. Díez-Pascual, J.A. Luceño Sánchez, R. Peña Capilla, P. García Díaz, Recent developments in graphene/polymer nanocomposites for application in polymer solar cells. *Polymers* **10**, 217 (2018). <https://doi.org/10.3390/polym10020217>
7. P. Singh, R. Saini, D. Deepika, R. Kumar, A. Singh, Novel composites: synergistic effects of Graphene Oxide, conducting polymers and metal oxides in supercapacitor electrodes. *J. Fluid Flow, Heat Mass Transf.* **11**, 322–334 (2024). <https://doi.org/10.11159/jffhmt.2024.032>
8. Y. Mao, L. Wang, C. Chen, Z. Yang, J. Wang, Thickness determination of ultrathin 2D materials empowered by machine learning algorithms. *Laser Photon. Rev.* (2023). <https://doi.org/10.1002/lpor.202200357>
9. G.K. Yogesh, D. Nandi, R. Yeetsorn, W. Wanchan, C. Devi, R.P. Singh, A. Vasistha, M. Kumar, P. Koinkar, K. Yadav, A machine learning approach for estimating supercapacitor performance of graphene oxide nano-ring based electrode materials. *Energy Adv.* **4**, 119–139 (2025). <https://doi.org/10.1039/D4YA00577E>
10. A. Barkhordari, H.R. Mashayekhi, P. Amiri, S. Özçelik, Ş Altındal, Y. Azizian-Kalandaragh, Machine learning approach for predicting electrical features of Schottky structures with graphene and ZnTiO₃ nanostructures doped in PVP interfacial layer. *Sci. Rep.* **13**, 13685 (2023). <https://doi.org/10.1038/s41598-023-41000-z>
11. B. Ersöz, A. Öter, Z. Berktaş, H.İ Bülbül, A. Di Bartolomeo, Ş Sağıroğlu, E. Orhan, Artificial intelligence-driven data generation for temperature-dependent current–voltage characteristics of diodes. *FlatChem.* **51**, 100847 (2025). <https://doi.org/10.1016/j.flatc.2025.100847>
12. R.O. Ocaya, A.A. Akinyelu, A.G. Al-Sehemi, A. Dere, A.A. Al-Ghamdi, F. Yakuphanoglu, Machine learning models for efficient characterization of Schottky barrier photodiode internal parameters. *Sci. Rep.* **13**, 13990 (2023). <https://doi.org/10.1038/s41598-023-41111-7>
13. J.Z. García-Córdova, J.A. Arano-Martinez, C. Mercado-Zúñiga, C.L. Martínez-González, C. Torres-Torres, Predicting the multiphotonic absorption in graphene by machine learning. *AI.* **5**, 2203–2217 (2024). <https://doi.org/10.3390/ai5040108>
14. A. Rabehi, B. Nail, H. Helal, A. Douara, A. Ziane, M. Amrani, B. Akkal, Z. Benamara, Optimal estimation of Schottky diode parameters using advanced swarm intelligence algorithms. *Semiconductors* **54**, 1398–1405 (2020). <https://doi.org/10.1134/S1063782620110214>
15. S. Papadopoulos, E. Azar, W.-L. Woon, C.E. Kontokosta, Evaluation of tree-based ensemble learning algorithms for building energy performance estimation. *J. Build. Perform. Simul.* **11**, 322–332 (2018). <https://doi.org/10.1080/19401493.2017.1354919>
16. M.A. Hassan, A. Khalil, S. Kaseb, M.A. Kassem, Exploring the potential of tree-based ensemble methods in solar radiation modeling. *Appl. Energy* **203**, 897–916 (2017). <https://doi.org/10.1016/j.apenergy.2017.06.104>
17. S. Kumar, K.S.K. Singh, K.K. Singh, Data-driven modeling for predicting tribo-performance of graphene-incorporated glass-fabric reinforced epoxy composites using machine learning algorithms. *Polym. Compos.* **43**, 6599–6610 (2022). <https://doi.org/10.1002/pc.26974>
18. P.K. Kanti, P. Paramasivam, V.V. Wanatasanappan, S. Dhanasekaran, P. Sharma, Experimental and explainable machine learning approach on thermal conductivity and viscosity of water based graphene oxide based mono and hybrid nanofluids. *Sci. Rep.* **14**, 30967 (2024). <https://doi.org/10.1038/s41598-024-81955-1>
19. S.B. Nadkarni, G.S. Vijay, R.C. Kamath, Comparative study of random forest and gradient boosting algorithms to predict airfoil self-noise, in *Proceedings of RAiSE-2023*. (MDPI, Basel Switzerland, 2023), p.24
20. K. Chahine, Tree-based algorithms and incremental feature optimization for fault detection and diagnosis in photovoltaic systems. *Eng.* **6**, 20 (2025). <https://doi.org/10.3390/eng6010020>
21. K. Prakash, A. James, N.J. Valeti, M.K. Singha, Optimization and numerical studies with machine learning assisted graphene-based CuSbS₂ thin film solar cell for flexible electronics applications. *J. Phys. Chem. Solids* **199**, 112513 (2025). <https://doi.org/10.1016/j.jpcs.2024.112513>
22. L. Breiman, Random forests. *Mach. Learn.* (2001). <https://doi.org/10.1023/A:1010933404324>

23. M. Schonlau, R.Y. Zou, The random forest algorithm for statistical learning. *Stata J.* **20**, 3–29 (2020). <https://doi.org/10.1177/1536867X20909688>
24. W. Tercha, S.A. Tadjer, F. Chekired, L. Canale, Machine Learning-based forecasting of temperature and solar irradiance for photovoltaic systems. *Energies* (2024). <https://doi.org/10.3390/en17051124>
25. J.S. Sohal, V. Hegde, B. Smitha, Solar power probabilistic prediction using random forest regressor, in *Renewable Energy Optimization Planning and Control: Proceedings of ICRTE-2022*. (Springer Nature, Singapore, 2023), pp.123–135
26. İ.F. Şener, İ Tuğal, Optimized CNN-LSTM with hybrid metaheuristic approaches for solar radiation forecasting. *Case Stud. Therm. Eng.* **72**, 106356 (2025). <https://doi.org/10.1016/j.csite.2025.106356>
27. C. Tianqi, H. Tong, xgboost: eXtreme Gradient Boosting. R Packag. (2018). <https://doi.org/10.1145/2939672.2939785>
28. J. Fan, X. Wang, L. Wu, H. Zhou, F. Zhang, X. Yu, X. Lu, Y. Xiang, Comparison of support vector machine and extreme gradient boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: a case study in China. *Energy Convers. Manag.* **164**, 102–111 (2018). <https://doi.org/10.1016/j.enconman.2018.02.087>
29. C. Bentéjac, A. Csörgő, G. Martínez-Muñoz, A comparative analysis of gradient boosting algorithms. *Artif. Intell. Rev.* **54**, 1937–1967 (2021). <https://doi.org/10.1007/s10462-020-09896-5>
30. H. Nhat-Duc, T. Van-Duc, Comparison of histogram-based gradient boosting classification machine, Random Forest, and deep convolutional neural network for pavement raveling severity classification. *Autom. Constr.* **148**, 104767 (2023). <https://doi.org/10.1016/j.autcon.2023.104767>
31. G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.Y. Liu, LightGBM: a highly efficient gradient boosting decision tree. *Adv. Neural. Inf. Process. Syst.* **30**, 3149–3157 (2017)
32. J. Park, J. Moon, S. Jung, E. Hwang, Multistep-ahead solar radiation forecasting scheme based on the light gradient boosting machine: a case study of Jeju Island. *Remote Sens.* (2020). <https://doi.org/10.3390/rs12142271>
33. S. Wang, Y. Ren, B. Xia, Estimation of urban AQI based on interpretable machine learning. *Environ. Sci. Pollut. Res.* **30**, 96562–96574 (2023). <https://doi.org/10.1007/s11356-023-29336-5>
34. J.L. Chen, Z.Z. Chen, Y.S. Chang, C.I. Li, T.I. Kao, Y.T. Lin, Y.Y. Xiao, J.F. Qiu, Intrusion detection system for AI box based on machine learning, in *International Conference on Advanced Communication Technology (ICACT)*. (IEEE, Republic of Korea, 2023), pp.111–116
35. M. Rahimi, M.R. Afrash, S. Shadnia, B. Mostafazadeh, P.E.T. Evini, M.S. Bardsiri, M. Ramezani, Prediction the prognosis of the poisoned patients undergoing hemodialysis using machine learning algorithms. *BMC Med. Inform. Decis. Mak.* (2024). <https://doi.org/10.1186/s12911-024-02443-0>
36. L. Oktovianus, K.B. Waluya, J. Surianto, A. Linardi, M.A. Sunusmo, E.B. Untoro, I.S. Edbert, Boosting algorithms in breast cancer classification: accuracy and performance metrics analysis. *IET Conf. Proc.* **2025**, 563–567 (2024). <https://doi.org/10.1049/icp.2025.0310>
37. R. Gayathri, S.U. Rani, L. Čepová, M. Rajesh, K. Kalita, A comparative analysis of machine learning models in prediction of mortar compressive strength. *Processes* **10**, 1387 (2022). <https://doi.org/10.3390/pr10071387>
38. S. Luo, T. Chen, Two derivative algorithms of gradient boosting decision tree for silicon content in blast furnace system prediction. *IEEE Access* **8**, 196112–196122 (2020). <https://doi.org/10.1109/ACCESS.2020.3034566>
39. P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees. *Mach. Learn.* **63**, 3–42 (2006). <https://doi.org/10.1007/s10994-006-6226-1>
40. U. Saeed, S.U. Jan, Y.-D. Lee, I. Koo, Fault diagnosis based on extremely randomized trees in wireless sensor networks. *Reliab. Eng. Syst. Saf.* **205**, 107284 (2021). <https://doi.org/10.1016/j.res.2020.107284>
41. Y. Lou, Y. Ye, Y. Yang, W. Zuo, G. Wang, M. Strong, S. Upadhyaya, C. Payne, Individualized empirical baselines for evaluating the energy performance of existing buildings. *Sci. Technol. Built Environ.* **29**, 19–33 (2023). <https://doi.org/10.1080/23744731.2022.2134680>
42. D. Chicco, M.J. Warrens, G. Jurman, The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Comput. Sci.* (2021). <https://doi.org/10.7717/PEERJ-CS.623>
43. İ Tuğal, Comparative analysis of LSTM architectures for wind speed forecasting: a case study in Muş, Turkey. *Turk. Dog. Fen Derg.* **13**, 107–119 (2024). <https://doi.org/10.46810/tdfd.1525648>
44. P.P. Phyto, Y.-C. Byun, Hybrid ensemble deep learning-based approach for time series energy prediction. *Symmetry* **13**, 1942 (2021). <https://doi.org/10.3390/sym13101942>
45. M.H. Alsharif, M.K. Younes, J. Kim, Time series ARIMA model for prediction of daily and monthly average global solar radiation: the case study of Seoul, South Korea. *Symmetry* **11**, 240 (2019). <https://doi.org/10.3390/sym11020240>

46. D.C. Montgomery, *Montgomery: Design and Analysis of Experiments* (John Wiley & Sons, Hoboken, NJ, USA, 2017)
47. M.H. Kutner, C.J. Nachtsheim, J. Neter, W. Li, Introduction to the design of experimental and observational studies, in *Applied Linear Statistical Models*. (McGraw-Hill/Irwin, Columbus, OH, USA, 2005)
48. A. Unwin, Discovering statistics using R by Andy Field, Jeremy Miles, Zoë Field. *Int. Stat. Rev.* (2013). https://doi.org/10.1111/insr.12011_21
49. J.W. Tukey, Comparing individual means in the analysis of variance. *Biometrics* (1949). <https://doi.org/10.2307/3001913>
50. S.E. Maxwell, H.D. Delaney, K. Kelley, *Designing Experiments and Analyzing Data: A Model Comparison Perspective*, 3rd Edition (2017). <https://doi.org/10.4324/9781315642956>
51. S. Kumar, P. Kaur, A. Gosain, A comprehensive survey on ensemble methods, in *2022 IEEE 7th International Conference on Convergence in Technology*. (IEEE, Pune, 2022), pp.1–7
52. V. Aceña, I. Martín de Diego, R.R. Fernández, J.M. Moguerza, Minimally overfitted learners: a general framework for ensemble learning. *Knowl. Based Syst.* **254**, 109669 (2022). <https://doi.org/10.1016/j.knosys.2022.109669>
53. E.I. Marchenko, S.A. Fateev, A.A. Petrov, V.V. Korolev, A. Mitrofanov, A.V. Petrov, E.A. Goodilin, A.B. Tarasov, Database of two-dimensional hybrid perovskite materials: open-access collection of crystal structures, band gaps, and atomic partial charges predicted by machine learning. *Chem. Mater.* **32**, 7383–7388 (2020). <https://doi.org/10.1021/acs.chemmater.0c02290>
54. L. Weston, C. Stampfl, Machine learning the band gap properties of kesterite I2–II–IV–V4 quaternary compounds for photovoltaics applications. *Phys. Rev. Mater.* **2**, 085407 (2018). <https://doi.org/10.1103/PhysRevMaterials.2.085407>
55. J. Leskovec, A. Rajaraman, J.D. Ullman, *Mining of Massive Datasets*, 2nd edition (2014). <https://doi.org/10.1017/CBO9781139924801>
56. Y. LeCun, Y. Bengio, G. Hinton, Deep learning. *Nature* **521**, 436–444 (2015). <https://doi.org/10.1038/nature14539>
57. T. Ko, T. Park, M. Kim, K. Min, Enhancing predictions of experimental band gap using machine learning and knowledge transfer. *Mater. Today Commun.* **41**, 110717 (2024). <https://doi.org/10.1016/j.mtcomm.2024.110717>
58. P. Xu, T. Lu, L. Ju, L. Tian, M. Li, W. Lu, Machine learning aided design of polymer with targeted band gap based on DFT computation. *J. Phys. Chem. B* **125**, 601–611 (2021). <https://doi.org/10.1021/acs.jpcc.0c08674>
59. Y. Zhang, C. Ling, A strategy to apply machine learning to small datasets in materials science. *Npj Comput. Mater.* **4**, 25 (2018). <https://doi.org/10.1038/s41524-018-0081-z>
60. C. Meng, S. Griesemer, D. Cao, S. Seo, Y. Liu, When physics meets machine learning: a survey of physics-informed machine learning. *Mach. Learn. Comput. Sci. Eng.* **1**, 20 (2025). <https://doi.org/10.1007/s44379-025-00016-0>
61. Q. Hua, M. Khazaei, A.A. Enkeshafi, M. Shen, S. Riahi, A. Rezaniakolaei, Physics-informed surrogate neural network for optimizing diode-based interfaces in leadless pacemaker energy harvesting. *IEEE Trans. Circuits Syst. II Express Briefs* **73**, 223–227 (2026). <https://doi.org/10.1109/TCSII.2025.3641701>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.