OPEN ACCESS

REGULAR ARTICLE

Intelligent Approach for Analyzing Semiconductor Band Gaps in Nanomaterial Systems

Bhagyashree Ashok Tingare[1], R.A. Kapgate[2], P. William[3], Jaikumar M. Patil[4], Tarun Dhar Diwan[5,*], Prasad M. Patare[6], Laxmikant S Dhamande[6]

[1] *Department of Artificial Intelligence and Data Science, D Y Patil College of Engineering, Akurdi, Pune*
[2] *Department of Mechatronics Engineering, Sanjivani College of Engineering, Kopargaon, MH, India*
[3] *Department of Information Technology, Sanjivani College of Engineering, Kopargaon, MH, India*
[4] *Department of Computer Science and Engineering, Shri Sant Gajanan Maharaj College of Engineering, Shegaon, SGBAU, Amravati*
[5] *Controller of Examination (COE), Atal Bihari Vajpayee University, Bilaspur, India*
[6] *Department of Mechanical Engineering, Sanjivani College of Engineering, Kopargaon, MH, India*

(Received 10 April 2025; revised manuscript received 15 August 2025; published online 29 August 2025)

Analysis of semiconductor band gaps in nanomaterials is of great importance for electronics applications. Traditional approaches have limitations in dealing with complex, nonlinear relationships for the prediction of band gaps. This study proposes a Fine-Tuned White Shark Algorithm-Resilient XGBoost (FWS-RXGBoost) model that eliminates the challenges associated with optimizing the hyperparameters of XGBoost for more robust predictions. A Kaggle dataset of material fingerprints and target band gap values are used. To ensure that the model accuracy, feature normalization by Z-score at preprocessing stage standardizes the features, which enhances the gradient-based learning. Optimization inspired by White Shark achieves a balance between the global exploration and local exploitation. This model is proven to be more resilient with noise in data. Comparisons have been made with a gradient boosting model and the Extra Trees model. According to RMSE (0.17), MAE (0.10), and R² score (0.97), FWS-RXGBoost is effective at modeling complex dependencies related to band gap predictions. In this regard, these results show that FWS-RXGBoost is a reliable, high-accuracy tool for the prediction of semiconductor band gaps and is presently ready for application in any real-world settings where accuracy is critical. Here, more varied datasets and sophisticated hybrid models may be used in future studies to increase prediction capabilities.

Keywords: Machine learning, Semiconductor band gaps, Nanomaterials, Fine-tuned White Shark algorithm-resilient XGBoost (FWS-RXGBoost).

DOI: 10.21272/jnep.17(4).04027 PACS numbers: 07.05.Mh, 73.20.At

1. INTRODUCTION

The analysis of the semiconductor band gap was crucial for further applications in energy storage, electronics, and optoelectronics since the working of materials was described by their optical and electrical properties (Terna et al., 2021). It was very challenging to predict band gaps in nanomaterials due to quantum effects at the nanoscale; hence, the traditional computational methods such as DFT are complex and computationally intensive. These shortcomings highlight the need for better band gap prediction techniques, especially as scientists look to identify and design new materials in an expedited manner (Burch et al., 2022). One promising alternative was machine learning, which relies on data-driven insights to predict properties based on material composition, structure, and other critical attributes. To discover such complex correlations between atomic and structural elements without requiring a lot of computing, an ML technique might be developed on a dataset of known band gap values for different nanomaterials (Prasad et al., 2022). It was a clever machine-learning framework that will be utilized in the future to expedite the production of semiconductors made of nanomaterials with high precision, rapidity, and expense effectiveness in band gap forecasting (Alcañiz et al., 2023). The creation of such models would assure improved next-generation semiconductor development, expedite material discovery procedures, optimize material design, and foster innovation in domains where more precise electronic characteristics are required. For use in electronics, photovoltaics, and other cutting-edge technologies, precise determination of nanomaterials' band gaps was essential in semiconductor research (Liu, et al., 2023). Although

[*] Correspondence e-mail: tarunctech@gmail.com

Cite this article as: B.A. Tingare et al., *J. Nano- Electron. Phys.* 17 No 4, 04027 (2025) https://doi.org/10.21272/jnep.17(4).04027

they work well, traditional approaches like Density Functional Theory (DFT) might be laborious and operationally taxing, particularly for intricate nanoscale systems where material characteristics are greatly impacted by quantum phenomena. The difficulty has led researchers to look for quicker and more scalable alternatives (Pandit et al., 2023). Nanomaterial properties and band gap were modelled by machine learning.

To enhance the accuracy in materials research, the current investigation seeks to apply and enhance advanced sophisticated machine learning models for prediction and analysis regarding semiconductor nanomaterial band gap properties.

An overview of relevant work is given in Part 2, and a technique is provided in Part 3. The performance evaluation is shown in Part 4, the discussion is shown in Part 5 and the conclusion is presented in Part 6.

2. RELATED WORK

The exertion was based on a novel extreme learning machine (ELM) computational intelligence method by utilizing the size of the compound's crystallite and lattice parameters as model characteristics to estimate Doped ZnSe's band gap energy nanostructured semiconductors by Aldhafferi et al., (2022). The created ELM-based model was compared with Support Vector Regression – Genetic Algorithm (SVR-GA) and Stochastic Partial Regression (SPR) models previously available in the fiction, using multiple performance indicators.

The performance of the established hybrid gravitational search (GS) centered multi-layer support vector regression model was compared with the traditional computational intelligence Support Vector Regression (SVR), Confidence Interval (CI) and stepwise regression (SWR) presented model in the literature in comparison with the advanced hybrid Grid Search-Machine Learning Support Vector Regression Model (GS-MLSVRM) with the existing SVRCI model and the Systematic Testing (ST)-based model in terms of the mean absolute percentage deviation measure (Shamasah et al., 2020).

An extreme learning machine and crystal distortion along with crystallite size were used by Souiyah et al. (2023) to use a strontium titanate magnetic photo-catalyst. With triangular basis (Tranba) sigmoid (Sig) activation functions, the established ELM-based models surpass the stepwise regression algorithm (SRA) model currently used in the fiction when measured using various presentation metrics, including the coefficient of correlation (CC), mean absolute error (MAE) and root mean square error (RMSE).

3. METHODOLOGY

The process goes through relevant datasets sourced from reliable sources. Data preparation, which involves cleaning and normalizing the data, results in consistency and dependability in the subsequent step. The suggested model will then be created utilizing cutting-edge machine-learning techniques. Lastly, appropriate performance measures like RMSE, MAE, and R^2 are used to evaluate the generated model's performance. Fig. 1 exemplifies the overall research flow.

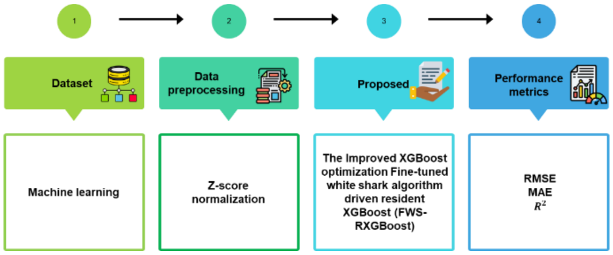
1
2
3
4
Dataset
Data preprocessing
Proposed
Performance metrics
Machine learning
Z-score normalization
The Improved XGBoost optimization Fine-tuned white shark algorithm driven resident XGBoost (FWS-RXGBoost)
RMSE
MAE
R^2

Fig. 1 – Overall research flow

3.1 Dataset

The subsequent section, "polymer", in the train and test data is the name of the polymer in simplified molecular-input line-entry system (SMILES) format. The 84 columns following the polymer name are material fingerprints that have been generated for the polymers. The target column is the "band_gap" column, which represents the band gap of the polymers**.**

3.2 Data Preprocessing Using Z-Score Normalization

The information is scaled using Z-score normalization, also known as standardization, which involves deducting the information's mean and in-between by the standard deviation. For algorithms that were sensitive to feature expanding, like gradient-based models, the procedure centers the data over a mean of 0 and sets the variance to 1. In the pre-processing step, normalization was a procedure that breaks down data into numerical properties and can transform values for data into a range of values. In data normalization, several techniques were often used, such as normalization by decimal scaling, and z-score normalization. Z-score normalization, as shown in Equation (1), converts a u_j value from attribute F to u into a previously unidentified range.

$$u' = \frac{u_j - F_j}{std(F)} \qquad (1)$$

Where u' = normalization value's outcome. u = the attribute's actual value to be adjusted. F_j= attribute's mean value. $std(F)$ = property $F$of the standard deviation.

3.3 Fine-tuned White Shark Algorithm-Driven Resilient XGboost (FWS-RXGboost)

FWS-RXGBoost employs sophisticated optimization techniques to increase the presentation of the RXGBoost model. The White Shark Algorithm, which is modeled after the white shark hunting strategy, optimally fine-tunes the XGBoost model's hyperparameters in a hybrid approach that balances exploration and exploitation in the pursuit of the best model configuration to employ to improve model robustness and make high-accuracy predictions. Resilience was given a major central role in this framework, indicating that this model will despite function rather well even with noisy data or supply distributions. The interaction between FWS and RXGBoost makes it easy to understand how the patterns have changed. This recently introduced combination produces quite good gains on classification tasks, making it a powerful tool in machine learning to deal with difficult data sets.

1.1.1 The Resilient XGBoost Algorithm

Through efficient hyperparameter tweaking and improved model resilience, the Resilient XGBoost Optimiser is used in this study to improve accuracy in forecasting for semiconductor band gaps. It is appropriate for the complicated nature of nanostructured material semiconductor characteristics because of its robustness, which enables it to continue operating even in the presence of noisy data or variable distributions. RXGBoost is a boosted tree model that aims to provide a more robust classifier model by integrating a large number of tree models. The gradient descent tree modification is also the method that has been applied to RXGBoost. The conventional Gradient-boosted decision trees (GBDT) approach merely makes use of XGBoost to perform a second-order Taylor on the first derivative. Making the loss function larger. This research presents a model of XGBoost for fusion multi-batch prediction. A multitasking learning process has been developed to learn characteristics from various batches. In the meantime, the prediction fusion process produces multi-feature fusion outcomes. The proposed RXGBoost model reduces gradient disappearance and model complexity by adding a regular term to the objective function, identifying the best solution free from overfitting.

Considering s samples and p feature wafer data sets, $C = \{(w_j, z_j)\}(|C| = t, w_j \in Q, z_j \in Q)$ the output result of L iterations is used by the boosted tree model. The anticipated price for the $j - th$ wafer, sample is z_j and it's the phrase is $\hat{z}_j$ is expressed in equation (1).

$$\hat{z}_j = \phi(w_j) = \sum_{l=1}^{L} e_l(w_j) \tag{2}$$

Calculations (2) and (3) demonstrate the loss function throughout the wafer yield forecasting model's instruction:

$$obj = \sum_j k(z_j, \hat{z}_j) + \sum_l \Omega\,(e_l) \tag{3}$$

$$\Omega(e_l) = \gamma^S + \frac{1}{2}\lambda\left||f_i|\right|^2 \tag{4}$$

The loss function is represented by $\sum_j k(z_j, \hat{z}_j)$, the regularization term is represented by $\sum_l \Omega\,(e_l)$, and the actual value of wafer yield is represented by z_j, and the anticipated amount of wafer yield is represented by $\hat{z}_j$.

A fresh regression tree is introduced to the model at a time throughout the model training process, which uses the gradient boosting technique to preserve the current models. Assume that in the $s - th$ iteration, the $j - th$ wafer sample's predicted outcome is $\hat{z}_j^{(s)}$. The newly added regression tree, $e_s(w_j)$ was derived in the manner described below:

$$\hat{z}_j^{(2)} = 0, e_1(w_j) = \hat{z}_j^{(0)} + e_1(w_j), \sum_{l=1}^{s} e_l(w_j) + e_2(w_j) = \hat{z}_j^{(1)} + e_2(w_j), \sum_{l=1}^{s} e_l(w_j) = \hat{z}_j^{(s-1)} + e_s(w_j) \tag{5}$$

It then changed (5) into (3) to get the formula (6) that follows.

$$obj^{(s)} = \sum_j k\ \left(z_j, \hat{z}_j^{(s-1)} + e_s(w_j)\right)\ + \Omega(e_l) + constant \tag{6}$$

Add a common term and do a second-order Taylor expansion of the goal variable.

$$obj^{(s)} \cong \sum_{j=1}^{t}\left[h_j e_s(w_j) + \frac{1}{2} g_j e_s^2(w_j)\right] + \Omega(e_l) = \sum_{j=1}^{t}\left[h_j\theta_{r(w)} + \frac{1}{2} g_j \theta_{r(w)}^2\right] + \gamma^S + \frac{1}{2}\lambda\|x_i\|^2 = \sum_{j=1}^{t}\left[\left(\sum_{j\in I_i} h_j\right)\theta_i + \frac{1}{2}\left(\sum_{j\in I_i} g_j + \lambda\right)f_i^2\right] + \gamma^S \tag{7}$$

Particularly in $h_j = \partial_{\hat{z}_j^{(s-1)}} k\left(z_j, \hat{z}_j^{(s-1)}\right), g_j = \partial^2 \hat{z}_j^{(s-1)} k\left(z_j, \hat{z}_j^{(s-1)}\right)$ Equation (7) may be simplified by defining (8) $H_j = \sum_{j\in I_i} h_i$, $G_j = \sum_{j\in I_i} g_j$ and substituting them into it,

$$obj^{(s)} = \sum_{i=1}^{S}\left[H_j\theta_i + \frac{1}{2}(G_j + \lambda)\theta_i^2\right] + \gamma^S \tag{8}$$

The value of the leaf node θ_i in equation (9) is unclear. As a result, the optimal value θ_i^* of leaf node j may be found by solving the objective function $obj^{(s)}$, which looks for the first derivative for θ_i^*.

$$\theta_i^* = -\frac{H_j}{G_j + \lambda} \tag{9}$$

When θ_i^* is substituted into the objective function, the minimal value is obtained by $obj^{(s)}$,

$$obj^{(s)} = -\frac{1}{2}\sum_{i=1}^{S}\frac{H_j}{G_j + \lambda} + \gamma^S \tag{10}$$

4. RESULT AND DISCUSSION

Findings from outcomes were conducted using the Python 3.11 software version. The research was perform using a laptop running Windows 10 with an Intel i7 CPU and 32 GB of RAM. The effectiveness of the aforementioned model is determined by comparison measures like RMSE, MAE, and R^2.The proposed method compared to other existing methods such as Gradient Boosting [17], Light Gradient Boosting Machine (LGBM) Regressor [17], and Extra Trees [17]. Better band gap robustness was indicated by the suggested models' more accurate band gap predictions and reduced error rate when compared to the existing models. In semiconductor band gap analysis, the main features of importance are particle size, crystal structure, and atomic composition. Atomic composition is the most significant factor, while importance ratings show the impact of characteristics. Figure 2 indicates the outcomes of the results.

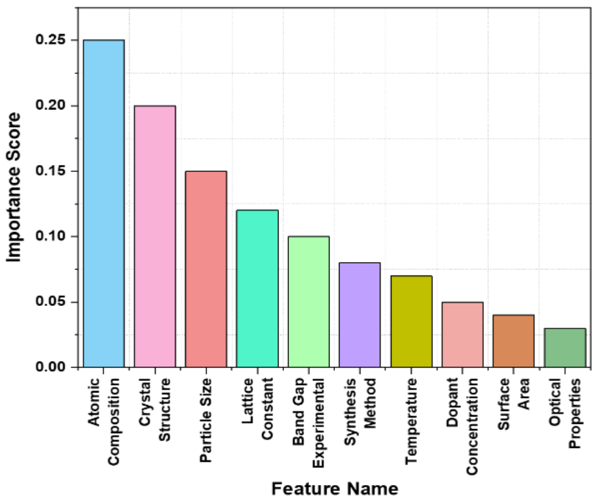

Fig. 2 – Outcomes of results

4.1 RMSE

A common metric in the assessment of the ordinary amount of the errors within predictions is root mean squared error. It indicates the square root of the average of the squared changes among the predictions and real values. The RMSE values, which indicate the number of mistakes in the models, decrease with the quality of the predictions made by the models. The suggested FWS-RXGboost model, as shown in Fig. 3 and Table 1, is superior to the current models when compared to gradient boosting, which has an RMSE of 0.18. In the end, this would suit real data better and forecast more accurately.

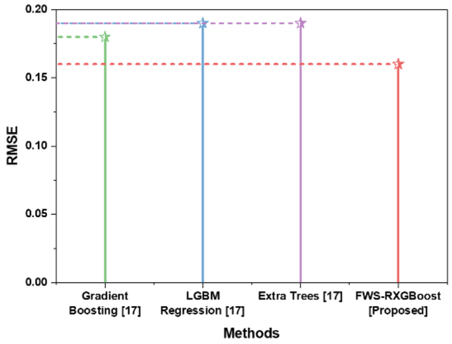

Fig. 3 – Result of RMSE

4.2 MAE

To measure the accuracy of predictions by the model is the average relative variation between actual and expected values, or Mean Absolute Error, or MAE. The lower the MAE, the more accurate the prediction. The MAE of LGBM and Gradient Boosting for the proposed model is 0.12. However, FWS-RXGboost has a higher MAE of 0.10; Figure 4 and Table 1 indicate better accuracy and predictive capacity.

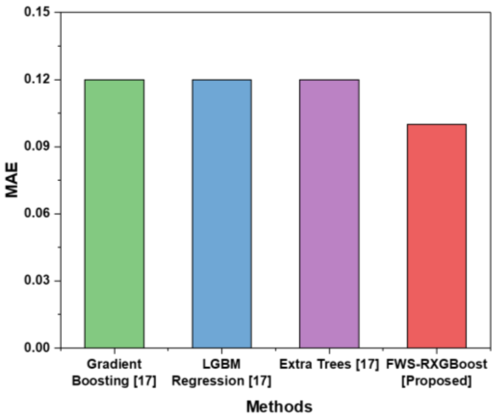

Fig. 4 – Result of MAE

1.2 R^2

The constant of drive measurement or R^2 is defined as the measure of the goodness of fit between model prediction and actual data. The larger the value for R^2, the better the predictability. The R^2 of the existing models of LGBM and Gradient Boosting is 0.96. The suggested model Figure 5 & Table 1. FWS-RXGboost has a greater value of $R^2 = 0.97$ hence, it provides greater predictability and is very close to the actual values.

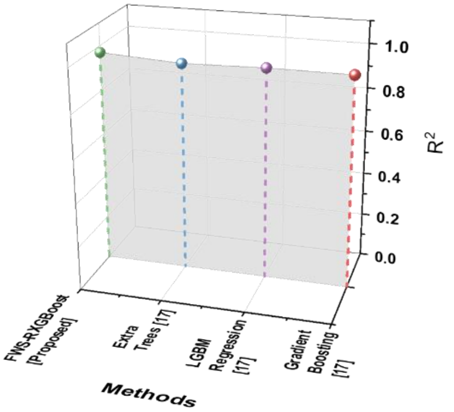

Fig. 5 – Result of R^2

Table 1 – Quantitative outcomes of the suggested methods

Method	**RMSE**	**MAE**	R^2
Gradient Boosting [17]	0.18	0.12	0.96
LGBM Regressor [17]	0.19	0.12	0.96
Extra Trees [17]	0.19	0.12	0.95
FWS-RXGboost [Proposed]	0.17	0.10	0.97

5. CONCLUSION

Band gaps in nanomaterials can be predicted and understood by combining cutting-edge machine learning techniques with an intelligent machine learning approach for analysing semiconductor band gaps in nanomaterial systems; the study successfully highlighted the significance of feature selection, data pre-processing, and gradient boosting optimization for precise outcomes using the FWS-RXGBoost model. The findings confirmed that machine learning techniques can accurately predict and assess semiconductor characteristics, dropping the essential for costly and time-consuming traditional experimental approaches. This produces strong prediction metrics, particularly when it comes to increases in RMSE, MAE, and R^2. Thus, these analyses have strengthened and solidified the idea that applying intelligent machine learning frameworks to gain an improved understanding of semiconductors could lead to improved discovery and testing of such materials for further advancements in materials science and nanotechnology, in addition to appreciating creative algorithms in material analysis.

REFERENCES

1. A.D. Terna, E.E. Elemike, J.I. Mbonu, O.E. Osafile, R.O. Ezeani, *Mater. Sci. Eng.: B* **272**, 115363 (2021).
2. M. Bursch, J.M. Mewes, A. Hansen, S. Grimme, *Angew. Chem. Int. Ed.* **61** No 42, e202205735 (2022).
3. K.R.K.V. Prasad, V.S. Rao, P. Harini, R.R. Mukiri, K. Ravindra, D.V. Kumar, R. Kasirajan, *J. Nanomater.* **2022**, 5450826 (2022).
4. A. Alcañiz, D. Grzebyk, H. Ziar, O. Isabella, *Energy Rep.* **9**, 447 (2023).
5. A.C. Liu, Y.Y. Lai, H.C. Chen, A.P. Chiu, H.C. Kuo, *Micromachines* **14** No 4, 764 (2023).
6. B. Pandit, S.R. Rondiya, S.F. Shaikh, M. Ubaidullah, R. Amaral, N.Y. Dzade, E.S. Goda, H.S. Gill, T. Ahmad, *J. Colloid Interface Sci.* **633**, 886 (2023).
7. J. Lu, F. Zhang, W.Y. Wang, G. Yao, X. Gao, Y. Liu, Z. Zhang, J. Wang, Y. Wang, X. Liang, H. Song, *J. Am. Ceram. Soc.* **106** No 11, 6923 (2023).
8. N. Aldhafferi, *Mater. Today Commun.* **31**, 103626 (2022).
9. S.M.I. Shamsah, T.O. Owolabi, *Chin. J. Phys.* **68**, 493 (2020).
10. A. Alqahtani, *J. Nanomater.* **2021** No 1, 4797686 (2021).
11. Y. Jia, X. Hou, Z. Wang, X. Hu, *ACS Sustain. Chem. Eng.* **9** No 18, 6130 (2021).
12. D. Hejazi, S. Liu, A. Farnoosh, S. Ostadabbas, S. Kar, *Machine Learning: Sci. Technol.* **1** No 2, 025007 (2020).
13. X. Chen, H. Lv, *NPG Asia Mater.* **14** No 1, 69 (2022).
14. M. Taniguchi, H. Takei, K. Tomiyasu, O. Sakamoto, N. Naono, *J. Phys. Chem. C* **126** No 29, 12197 (2022).
15. M. Souiyah, *Cogent Eng.* **10** No 1, 2232596 (2023).
16. M.H. Zeb, A. Rehman, M. Siddiqah, Q. Bao, B. Shabbir, M.Z. Kabir, *Adv. Theory Simul.* **7** No 7, 2400190 (2024).

Інтелектуальний підхід до аналізу заборонених зон у напівпровідникових наноматеріалах

Bhagyashree Ashok Tingare[1], R.A. Kapgate[2], P. William[3], Jaikumar M. Patil[4], Tarun Dhar Diwan[5],
Prasad M. Patare[6], Laxmikant S Dhamande[6]

[1] *Department of Artificial Intelligence and Data Science, D Y Patil College of Engineering, Akurdi, Pune*
[2] *Department of Mechatronics Engineering, Sanjivani College of Engineering, Kopargaon, MH, India*
[3] *Department of Information Technology, Sanjivani College of Engineering, Kopargaon, MH, India*
[4] *Department of Computer Science and Engineering, Shri Sant Gajanan Maharaj College of Engineering, Shegaon, SGBAU, Amravati*
[5] *Controller of Examination (COE), Atal Bihari Vajpayee University, Bilaspur, India*
[6] *Department of Mechanical Engineering, Sanjivani College of Engineering, Kopargaon, MH, India*

Аналіз заборонених зон у напівпровідникових наноматеріалах має велике значення для застосування в електроніці. Традиційні підходи мають обмеження в роботі зі складними, нелінійними зв'язками для прогнозування заборонених зон. У роботі пропонується модель FWS-RXGBoost (Fine-Tuned White Shark Algorithm-Resilient XGBoost), яка усуває проблеми, пов'язані з оптимізацією гіперпараметрів XGBoost для більш надійних прогнозів. Використовується набір даних Kaggle про відбитки матеріалів та цільові значення забороненої зони. Для забезпечення точності моделі нормалізація ознак за Z-оцінкою на етапі попередньої обробки стандартизує ознаки, що покращує градієнтне навчання. Оптимізація, натхненна White Shark, досягає балансу між глобальним дослідженням та локальним використанням. Ця модель виявилася більш стійкою до шуму в даних. Були проведені порівняння з моделлю градієнтного бустингу та моделлю Extra Trees. Згідно з показниками RMSE (0,17), MAE (0,10) та R² (0,97), FWS-RXGBoost ефективний для моделювання складних залежностей, пов'язаних з прогнозами забороненої зони. У зв'язку з цим, ці результати показують, що FWS-RXGBoost є надійним, високоточним інструментом для прогнозування ширини заборонених зон напівпровідників і наразі готовий до застосування в будь-яких реальних умовах, де точність є критично важливою. У майбутніх дослідженнях можуть бути використані більш різноманітні набори даних та складні гібридні моделі для розширення можливостей прогнозування.

Ключові слова: Машинне навчання, заборонені зони у напівпровіднику, наноматеріали, алгоритм White Shark, стійкий до XGBoost (FWS-RXGBoost).