PHYSICAL SCIENCES

MACHINE LEARNING ASSISTED ANALYSIS OF THE REFRACTIVE INDEX–BAND GAP RELATIONSHIP IN SEMICONDUCTOR MATERIALS

Mammadova V.
Ph.D. in Physics, Lecturer, Department of Optics and Molecular Physics, Baku State University, Baku State University
https://doi.org/10.5281/zenodo.21649096

Abstract

In this article, the relationship between the band gap width $E_g$and the refractive index nin semiconductor materials was investigated using a data-driven approach. Linear regression and Random Forest models were constructed based on literature data for II–VI and III–V group materials. The comparison showed that the Random Forest model takes local variations into account better and provides a closer prediction for ZnO. This approach is useful for the preliminary evaluation of optical properties.

Keywords: Semiconductors, Band gap energy, Refractive index, Machine learning, ZnO.

PACS: 07.05.Kf; 07.05.Mh; 78.20.Ci; 78.40.Fy

1. Introduction

The refractive index (n) is one of the key optical parameters that characterizes the interaction of a material with light and is defined as the ratio of the speed of light in vacuum to its propagation speed in the material. Also, the refractive index is a quantity that determines how much the path of radiation is bent in a substance. The refractive index depends on the optical absorption of the substance. Materials with a high absorption coefficient generally have a higher refractive index, whereas materials with a low absorption coefficient usually have a lower refractive index. As is known, optical absorption is mainly related to electronic transitions from the valence band to the conduction band, and the energy threshold of these transitions is determined by the band gap width (Eg). From this point of view, the relationship between the refractive index $(n)$and the optical band gap width has emerged [1].

In recent years, interest has increased in analyzing the relationship between the refractive index and the band gap using both empirical and dataset approaches. There are several reasons for this. The first reason is that the relationship between these two parameters is important from the point of view of evaluating the material's suitability for transparency, absorption, dispersion, and optoelectronic applications. The second reason is that the classical empirical models, such as Ravindra [2], Moss [3], Herve–Vandamme [4], Reddy [5], and Kumar–Singh [6], although they try to explain the $E_g \sim n$ relationship, do not work with the same accuracy for all materials. The constants used during the application of these models often depend on the chemical composition, whether the transition is direct or indirect, the structure, and the material group. Although these models appear simple in form, they are not universally applicable in practice. Therefore, a more detailed analysis of this dependence is needed. [7]. The third reason is the development of the machine learning approach. Thus, although the $Eg{\sim}n$ relationship has previously been evaluated using simple empirical formulas, it is now possible to investigate this relationship based on DFT calculations and machine learning approaches. These approaches make it possible to obtain results in a shorter time, and in some cases within seconds. İn modern studies, the band gap width is used to predict the refractive index (n) [8]. The most important reason is related to practical applications. Therefore, suitable values of both (E_g) and (n) are required for photonics, solar cells, sensors, optical coatings, transparent conductive oxides, and semiconductors with high refractive indices [1].

This allows us to say that it is not enough for a material to have only a large or small band gap width, it is also important to know how it interacts with light. In recent studies, an initial screening is first performed based on machine learning to identify semiconductors with high refractive indices, and then the results are verified using first-principles calculations.

In our previous theoretical analysis, it was shown that an inverse dependence exists between the refractive index (n) and the band gap energy (E_g) in semiconductor materials. However, this dependence is not completely linear, and the accuracy of the models varies depending on the material and the energy range [9]. Machine learning approaches make it possible to learn the dependence directly from the dataset without relying on a previously defined functional form and, as a result, to reveal hidden, nonlinear regularities [10–12].

The aim of this study is to analyze the relationship between the band gap energy E_g and the refractive index n in semiconductor materials based on a dataset of experimental values taken from the literature, and to comparatively investigate this relationship using linear regression and Random Forest models. The dataset used is based on literature data on II–VI and III–V semiconductors presented by Ahmad and co-authors [13]. Using this dataset, it was possible to compare the results obtained through the machine learning approach with experimental data and evaluate the extent to which the predictions differ from those data.

ZnO was selected in order to show the practical application possibilities of the models constructed on

the basis of the machine learning approach. The prediction of the refractive index (n) for ZnO was carried out. The main purpose of selecting ZnO is that it has a number of advantages and a wide field of application. ZnO is a wide-bandgap semiconductor and has a band gap width of approximately 3,37 eV at room temperature, and it is widely used in optoelectronic applications [14].

2. Methodology

As noted, for different semiconductor materials, the band gap energy E_g covers a wide energy range $(0{,}17 \div 6{,}2 \text{ eV})$, and the corresponding refractive index $n$values are presented in tabular form [13] (Table 1). Using the machine learning approach, a comparative analysis was carried out based on the dataset given in table [15, 16].

Machine learning is one of the main fields of artificial intelligence and is a computational approach used for training models and making predictions based on data. As we know, machine learning is divided into three types: supervised learning, unsupervised learning, and reinforcement learning methods.

Supervised machine learning is most suitable for neural networks that need to solve classification problems and works based on a large amount of input data.

Unsupervised learning is, learning without a teacher. This mainly includes the transmission of unlabeled data and trying to find common features and relationships within them. We can understand this as follows: neural networks are first trained on broad and diverse data and then applied to the solution of real-world tasks. The use of unlabeled data allows the model to reveal regularities in new examples and to generalize better [17].

In the reinforcement learning method, it includes the neural network receiving input data and processing it randomly. Then its output is evaluated based on certain criteria.

In this study, machine learning was adopted as a general methodological approach, and two supervised learning models Linear Regression and Random Forest were used for prediction.

Linear regression is a regression algorithm. İt means that the purpose is to predict a continuous numerical value. The Random Forest model, on the other hand, is one of the supervised machine learning algorithms and is based on the joint operation of many decision trees. Since this approach is not based on previously determined empirical formulas, it allows a freer analysis of the $E_g \sim n$ relationship based on data.

All calculations were carried out in the Python programming environment, and the scikit-learn library was used in the construction of the models [15].

The dependence between the band gap width (E_g) and the refractive index (n) was analyzed for the semiconductor materials given in Table 1. Based on experimental literature data, a program was written in the Python programming language, and the calculations were carried out in the Anaconda environment.

3. Results and Discussion

Each point in the obtained graph corresponds to a specific material and was plotted using the (E_g) and (n) values given for that material (Figure 1). The circular points in the figure show the experimental values taken from the literature, while the straight blue line shows the regression line obtained because of linear regression.

The regression line expresses the general linear tendency between (E_g) and (n). However, since this approach evaluates all points within the same simple linear dependence, it cannot fully explain the values that deviate from the linear trend. The main purpose of using the Random Forest model is to take such local differences and nonlinear changes into account better. Apparently, as (E_g) increases the refractive index (n) decreases. However, the observed dependence does not have a completely linear character, and local differences exist in different material groups.

Table 1.

Parameters of the semiconductor materials used by us.

№	Material	Semiconductor group	Band gap energy, Eg (eV)	Refractive Index n
1.	InSb	III-V	0.17	3.96
2.	InAs	III-V	0.36	3.5
3.	GaSb	III-V	0.73	3.8
4.	InN	III-V	0.7	2.9
5.	InP	III-V	1.34	3.1
6.	GaAs	III-V	1.43	3.3
7.	AlSb	III-V	1.62	2.9
8.	CdTe	II-VI	1.5	2.8
9.	CdSe	II-VI	1.74	2.7
10.	AlAs	III-V	2.16	2.9
11.	GaP	III-V	2.26	3.1
12.	ZnTe	II-VI	2.26	2.7
13.	CdS	II-VI	2.42	2.5
14.	ZnSe	II-VI	2.7	2.5
15.	ZnO	II-VI	3.37	2

16.	GaN	III-V	3.4	2.3
17.	ZnS	II-VI	3.68	2.35
18.	AlN	III-V	6.2	2.1

Therefore, this graph should not be evaluated as a universal law for all semiconductors, but as an initial analysis showing the general tendency within the used data. This model gives results closer to the real points especially in the medium and high (E_g) interval, which shows the advantage of the machine learning approach [10].

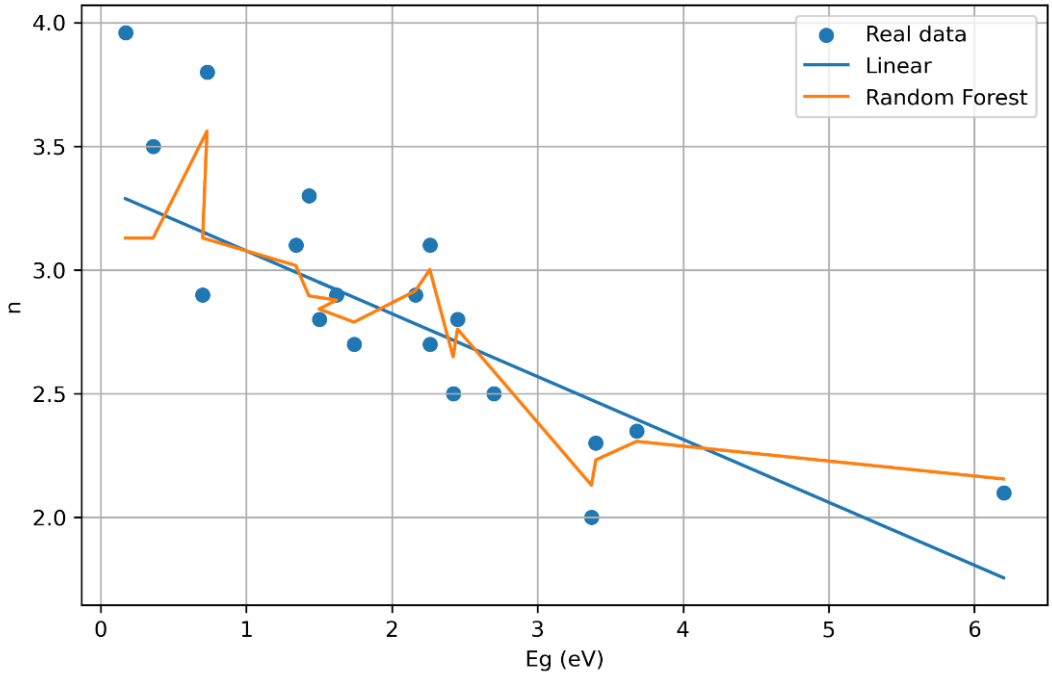

Figure 1. $\boldsymbol{E_g \sim n}$ *dependence on semiconductor materials*

It should be noted that since the literature dataset used is limited, it may cause certain fluctuations in the Random Forest model. Using a larger dataset may allow this problem to be reduced.

Although we carried out the prediction for ($N = 18$) semiconductors, we selected ZnO for specificity. Clearly, from Table 1, for ZnO, at the value $E_g = 3{,}37$ eV, the refractive index is equal to $n = 2$.

Based on the results, it was determined that values of $n \approx 2{,}468$ for the linear regression model and $n \approx 2{,}118$ for the Random Forest model were obtained (Table 2).

Table 2.

ML prediction of the refractive index (n)for ZnO.

Material	*Eg (eV)*	*n (LR)*	*n (RF)*
ZnO	**3,37**	**2,468**	**2,117**

The comparison of the results obtained for ZnO shows that the Random Forest model gives a result closer to the experimental and literature values than the Linear Regression model. In the dataset used, the value $n = 2{,}00$ is given for ZnO. The Linear Regression model predicted this value as $n = 2{,}468$, giving a relatively high value. The Random Forest model, however, gave the result $n = 2{,}117$, which is closer to the known experimental range for ZnO. This difference can be explained by the fact that the Linear Regression model follows the general linear tendency but cannot take local nonlinear changes into account. The Random Forest model, however, learned the local features in the data more quickly and gave a more suitable result for this specific case. Nevertheless, taking into account the limited size of the data set, the RF result was evaluated not as a final experimental value, but as an initial prediction result.

To evaluate the reliability of the model we used, 5-fold cross-validation was applied. In this method, the database is divided into five parts; at each stage, four parts are used for training the model, and one part is used for validation. The process is repeated five times, and the final result is determined based on the average value over all stages. This allows us to evaluate the stability and prediction accuracy of the model more reliably. For the evaluation of regression models, the mean squared error indicator — MSE — was also used. The average indicator of the MSE values obtained in the cross-validation stages was accepted as CV-MSE [18, 19].

The Random Forest model gave better results than the linear regression model (Train MSE: 0,0835; CV-MSE: $0{,}2141 \pm 0{,}2067$)in terms of both train MSE $(0{,}0217)$(train MSE shows how much error the model makes on the data it has learned) and CV-MSE $(0{,}1250 \pm 0{,}1222)$(the smaller the CV-MSE, the better the generalization ability of the model is considered) indicators (Table 3). This result quantitatively confirms that the Random Forest model learns the nonlinear relationships in the data more accurately. The difference between the train MSE $(0{,}0217)$and CV-MSE $(0{,}1250)$of the Random Forest model indicates that the model shows a certain degree of overfitting tendency, that is, it gives results that are excessively close to the values in the dataset. This case is mainly explained by the limited size of the dataset used $(N = 18)$. İn small datasets, complex models may show a tendency to memorize the data instead of generalizing it.

Table 3.

Train MSE and CV-MSE results for Linear Regression and Random Forest models.

Model	*Train MSE*	*CV-MSE*
Linear reqression	**0,0835**	**0,2141±0,2067**
Random Forest	**0,0217**	**0,1250±0,1222**

4. Conclusion

In this study, the (E_g) and (n) values given for II–VI and III–V semiconductor materials were used to analyze the $E_g \sim n$ dependence through Linear Regression and Random Forest models. The conducted analyses showed that as (E_g) increases, the refractive index (n) decreases. However, this dependence does not have a completely linear character for all materials, and local differences are observed in separate material groups.

Although the linear regression model is useful in terms of showing the general direction of the $Eg \sim n$ relationship, this model cannot fully reflect the nonlinear changes and local differences in the data. The Random Forest model, however, gave more suitable results because it takes into account the more complex and nonlinear dependences existing in the data. This is especially clearly seen in the prediction carried out for ZnO. While the refractive index for ZnO is given in the literature as $n = 2{,}0$, the linear regression model predicted this value as $n = 2{,}468$, and the Random Forest model predicted it as $n = 2{,}117$. As can be seen, the Random Forest model gave a result closer to the known value for ZnO.

To evaluate the reliability of the model, 5-fold cross-validation was applied, and Train MSE and CV-MSE indicators were compared. The fact that the Train MSE and CV-MSE values for the Random Forest model are lower compared with the linear regression model shows that this model learns the nonlinear relationships in the data better. At the same time, the difference between Train MSE and CV-MSE in the Random Forest model shows that a certain overfitting tendency may exist. This case can mainly be associated with the limited size of the dataset used.

Thus, the conducted study shows that the machine learning approach is a useful and applicable approach for the preliminary evaluation of the optical properties of semiconductor materials. In particular, the Random Forest model makes it possible to analyze the $Eg \sim n$ dependence not only as a simple linear tendency, but in a more complex and nonlinear form. However, the obtained results should be accepted not as a final experimental value, but as an initial prediction and comparative evaluation. In future studies, the use of a wider database, additional structural and optical parameters may make it possible to increase the accuracy and generalization ability of the model.

The application details of the machine learning approach are given in Appendix A.

References:

1. Gomaa H. M., Yahia I. S., Zahran H. Y. Correlation between the static refractive index and the optical bandgap: Review and new empirical approach // Physica B: Condensed Matter. 2021. Vol. 620. Article 413246. DOI: 10.1016/j.physb.2021.413246.

2. Ravindra N. M., Ganapathy P., Choi J. Energy gap–refractive index relations in semiconductors: An overview // Infrared Physics & Technology. 2007. Vol. 50. P. 21–29.

3. Moss T. S. A relationship between the refractive index and the infrared threshold of sensitivity for photoconductors // Proceedings of the Physical Society B. 1950. Vol. 63. P. 167.

4. Herve P., Vandamme L. K. J. General relation between refractive index and energy gap in semiconductors // Infrared Physics & Technology. 1994. Vol. 35. P. 609–615.

5. Reddy R. R., Ahammed Y. N., Azeem P. A., Narasimhulu K. V. Empirical relationship between refractive index and band gap of semiconductors // Journal of Materials Science. 2006. Vol. 41. P. 3071–3074.

6. Kumar A., Singh K. Estimation of refractive index from energy gap // Optik. 2012. Vol. 123. P. 496–499.

7. Lamichhane A. Energy-Gap-Refractive Index Relations in Semiconductors Using Wemple–DiDomenico Model to Unify Moss, Ravindra, and Herve–Vandamme Relationships // Solids. 2023. Vol. 4, No. 4. P. 316–326. DOI: 10.3390/solids4040020.

8. Einabadi E., Mashkoori M. Predicting refractive index of inorganic compounds using machine learning // Scientific Reports. 2024. Vol. 14. Article 24204.

9. Mammadova V. Theoretical Analysis of the Variation of Refractive Index with Band Gap in Semiconductor Materials Using Different Models // Norwegian Journal of Development of the International Science. 2025. No. 164. P. 58–66. DOI: 10.5281/zenodo.17121604.

10. Pilania G., Wang C., Jiang X., Rajasekaran S., Ramprasad R. Accelerating materials property predictions using machine learning // Scientific Reports. 2013. Vol. 3. Article 2810.

11. Ward L., Agrawal A., Choudhary A., Wolverton C. A general-purpose machine learning framework for predicting properties of inorganic materials // npj Computational Materials. 2016. Vol. 2. Article 16028.

12. Butler K. T., Davies D. W., Cartwright H., Isayev O., Walsh A. Machine learning for molecular and materials science // Nature. 2018. Vol. 559. P. 547–555.

13. Ahmad S., Ashraf M., Ahmad A., Singh D. V. Electronic and Optical Properties of Semiconductor and Alkali Halides // Arabian Journal for Science and Engineering. 2013. Vol. 38. P. 1889–1894. DOI: 10.1007/s13369-013-0551-z.

14. Özgür Ü., Alivov Y. I., Liu C., Teke A., Reshchikov M. A., Doğan S., Avrutin V., Cho S. J., Morkoç H. A comprehensive review of ZnO materials and devices // Journal of Applied Physics. 2005. Vol. 98. Article 041301.

15. Pedregosa F., Varoquaux G., Gramfort A. et al. Scikit-learn: Machine Learning in Python // Journal of Machine Learning Research. 2011. Vol. 12. P. 2825–2830.

16. Mammadova V. J. Machine learning in physics and semiconductor materials: approaches to modeling and prediction // Space Science Journal. DOI: 10.30546/209501.101.2026.3.01.316.

17. Bishop C. M. Pattern Recognition and Machine Learning. Springer, 2006.

18. Kohavi R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection // Proceedings of the 14th International Joint Conference on Artificial Intelligence. 1995. P. 1137–1143.

19. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. Springer, 2009.

Appendix A. Machine Learning Implementation

This appendix provides the Python implementation used to construct the machine learning models and to predict the refractive index for the ZnO.

```
# Machine Learning implementation for Eg~ n prediction
import pandas as pd
from sklearn.linear_model import LinearRegression
from sklearn.ensemble import RandomForestRegressor
# Dataset
data = {
 "Eg": [0.17, 0.36, 0.73, 0.70, 1.34, 1.43, 1.62, 1.50, 1.74,
 2.16, 2.26, 2.26, 2.42, 2.45, 2.70, 3.37, 3.40, 3.68, 6.20],
 "n": [3.96, 3.50, 3.80, 2.90, 3.10, 3.30, 2.90, 2.80, 2.70,
 2.90, 3.10, 2.70, 2.50, 2.80, 2.50, 2.00, 2.30, 2.35, 2.10]
}
df = pd.DataFrame(data)
# Feature and target
X = df["Eg"]]
y = df["n"]
# Model construction
lr = LinearRegression()
lr.fit(X, y)
rf = RandomForestRegressor(n_estimators=100, random_state=42)
rf.fit(X, y)
# Prediction for ZnO photodetector
X = df[["Eg"]]
zno_case = pd.DataFrame({"Eg": [3.37])
lr_pred = lr.predict(zno_case)[0]
rf_pred = rf.predict(zno_case)[0]
# Prediction table
result = pd.DataFrame({
 "Material": ["ZnO photodetector"],
 "Band Gap Energy Eg (eV)": [3.37],
 "Predicted n (Linear Regression)": [lr_pred],
 "Predicted n (Random Forest)": [rf_pred]
})
print(result.round(4))
```