Machine Learning-based Prediction of Optical Band Gaps in ZnO Thin Films

Ferhat Uçar and Nida Katı

Electrical–Electronics Engineering Department, Fırat University
Metallurgical and Materials Engineering Department, Fırat University

7th June 2023

Outline

1. Introduction
   - What is data science?
   - What do we do propose here?
2. Modeling and Analysis
   - ZnO thin film dataset
   - Machine Learning Algorithm–Random Forest Model
   - Analysis and Results
3. Discussions

Meaning of data science

We live in "data era" so,

What does data science mean?

- For a biologist, examining DNA sequences.
- For a banker, predicting the stock market.
- For a materials scientist, modeling the structure of complex materials.
- For a machine learning scientist, models and algorithms.

- Yet, one thing that is constant in all those different meanings.
- The concept of uncertainty.
- We learn from the data.
- Because the latent information is still in the data.
- It is unprocessed and waiting to be analysed...

Meaning of data science

We live in "data era" so,

What does data science mean?

- For a biologist, examining DNA sequences.
- For a banker, predicting the stock market.
- For a materials scientist, modeling the structure of complex materials.
- For a machine learning scientist, models and algorithms.

- Yet, one thing that is constant in all those different meanings.
- The concept of uncertainty.
- We learn from the data.
- Because the latent information is still in the data.
- It is unprocessed and waiting to be analysed...

Meaning of data science

We live in "data era" so,

What does data science mean?

- For a biologist, examining DNA sequences.
- For a banker, predicting the stock market.
- For a materials scientist, modeling the structure of complex materials.
- For a machine learning scientist, models and algorithms.

- Yet, one thing that is constant in all those different meanings.
- The concept of uncertainty.
- We learn from the data.
- Because the latent information is still in the data.
- It is unprocessed and waiting to be analysed...

Motivation

To produce a model for the prediction of the Energy bandgap (E_g) of ZnO thin films.

ZnO thin films

With its nanostructured design, ZnO thin film is of great interest in a variety of applications, devices, and sensors used in the defense industry, photonics, spintronics, and optoelectronics.

Why the prediction of the E_g? Why do we need our model?

Adapting the E_g has a key role in ZnO thin films. The specified process parameters, the determining of dopant, and various design forms of these two have important and decisive impacts on the microstructure and crystal structure of the produced films, and therefore on their optical performance.

Motivation

To produce a model for the prediction of the Energy bandgap (E_g) of ZnO thin films.

ZnO thin films

With its nanostructured design, ZnO thin film is of great interest in a variety of applications, devices, and sensors used in the defense industry, photonics, spintronics, and optoelectronics.

Why the prediction of the E_g? Why do we need our model?

Adapting the E_g has a key role in ZnO thin films. The specified process parameters, the determining of dopant, and various design forms of these two have important and decisive impacts on the microstructure and crystal structure of the produced films, and therefore on their optical performance.

The proposed model

In this study, the random forest (RF) machine learning (ML) algorithm which is a widespread usage in fitting and regression processing is designed to explain statistical correlation between the grain size, and lattice parameters along with E_g for various type of ZnO thin films.

Contributions

- To analyze a fitting model to investigate and clarify the correlations in ZnO production.
- To search for a regression model which could give a benefit to production process of ZnO thin films.

The proposed model

In this study, the random forest (RF) machine learning (ML) algorithm which is a widespread usage in fitting and regression processing is designed to explain statistical correlation between the grain size, and lattice parameters along with E_g for various type of ZnO thin films.

Contributions

- To analyze a fitting model to investigate and clarify the correlations in ZnO production.
- To search for a regression model which could give a benefit to production process of ZnO thin films.

The dataset definition

- The public dataset used in the experiments is obtained from (Zhang et al., 2020).
- The dataset includes a wide array of ZnO thin films that are fabricated through different synthesis ways and doped with numerous metal elements. The lattice parameters, and measured grain size are used as descriptors for revealing the correlation with E_g.

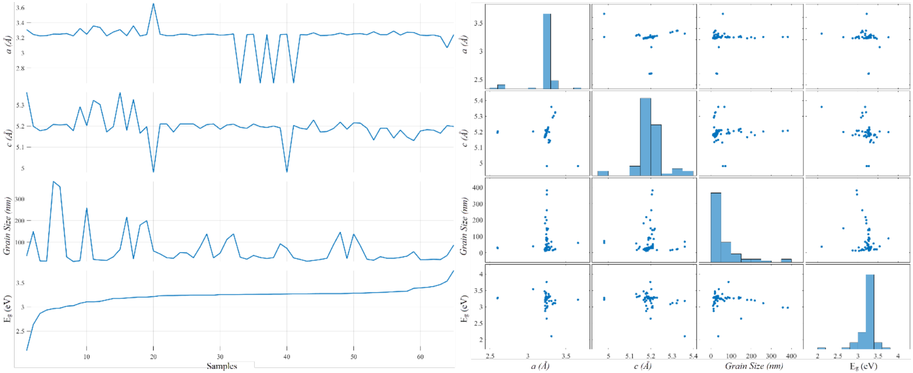


Figure: Dataset distribution graphics

The dataset definition

Figure: Dataset distribution graphics

Table 1: Some Important Statistical Measures of Dataset

Statistics	*a* (Å)	*c* (Å)	Grain size (nm)	Eg (eV)
Count	65.000000	65.000000	65.000000	65.000000
Mean	3.217916	5.196658	65.987385	3.220138
Standard Deviation	0.171353	0.059278	77.352870	0.208404
Minimum	2.598000	4.980000	11.000000	2.100000
25%	3.232700	5.180200	21.440000	3.200000
50%	3.244200	5.194000	31.050000	3.262000
75%	3.253000	5.210000	79.240000	3.283000
Maximum	3.660000	5.360000	382.000000	3.760000

Random Forest (RF)

- RF is a bagging-based machine learning model, which is widely been implicated as an ensemble learning algorithm to give a solution for various types of regression problems.
- The RF algorithm utilizes the classification and regression trees (CART) along with the ensemble methods (boosting or bagging method) to enhance the efficiency of the CART algorithm.
- RF model itself includes a number of decision trees as being less parameter containing.
- RF algorithm tries to reach the uncorrelated forest of decision trees whose overall accuracy is higher than other individual trees.

Analysis

Experiments

- In our study, we have two different types of experiment approaches, one is the fitting logic and the other one is regression logic.
- In both experiments, we use the RF model parameters as maxfeatures=3, numberofestimators=5000.
- The code design of the study is programmed on a workstation with Linux operating system using the Python Scikit-learn ML package and Jupyter Lab.

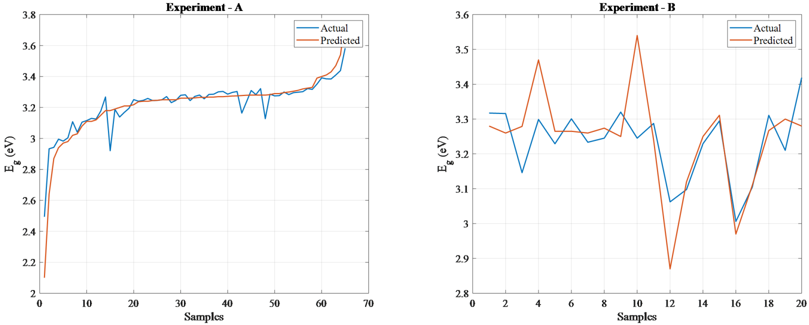
Analysis

Experiments

- The well-known performance metrics related to our models are R-squared, explained variance score (EVS), mean absolute percentage error (MAPE), mean square error (MSE), and root mean square error (RMSE).
- Experiment – A is for the fitting logic and Experiment – B is for the regression model logic while the former uses all the samples.
- The latter uses a train–test split in the 80% - 20% range.

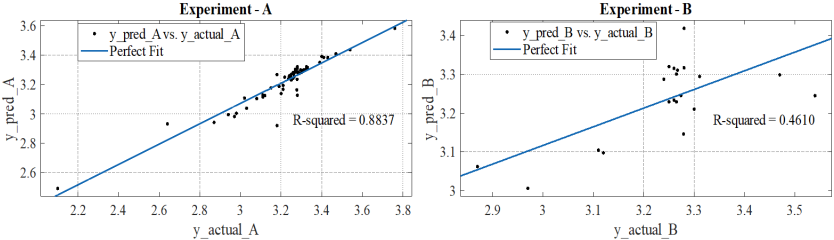
Results

Figure: Predicted vs. Actual tracing plots



Results

Figure: Correlation plots



Results

Table 2: Experimental results – Performance Values

Metrics	Experiment-A	Experiment-B
R-squared	0.8837	0.4610
EVS	0.8419	0.4601
MAPE	0.0146	0.0229
MSE	0.0067	0.0109
RMSE	0.0823	0.1045

Conclusion

- In our study, ML-based models are designed for the estimation of the E_g values of various thin ZnO thin films.
- Obtain results show that Experiment-A can correctly fit the distribution while Experiment-B deals with a more challenging task.
- Results indicate that our fitting model generates a high performance while the regression model is forced by non-uniform distribution.

Future work

Soon...

Future works cover building our own-built dataset of different nanostructured materials to make enhanced ML prediction models for physical and optical features.

Thanks for listening!

For your precious comments, please kindly rich me @ fucar@firat.edu.tr