DOI: 10.63527/1607-8829-2023-2-59-71

Anatychuk L. I.

L. I. Anatychuk, *acad. National Academy of Sciences of Ukraine*[1,2]
M. M. Korop [1]

[1] Institute of Thermoelectricity of the NAS and MES of Ukraine,
1 Nauky str., Chernivtsi, 58029, Ukraine;
e-mail: anatych@gmail.com, mykola.korop@chnu.edu.ua
[2] Yuriy Fedkovych Chernivtsi National University, 2 Kotsiubynskyi str., Chernivtsi, 58000, Ukraine
e-mail: anatych@gmail.com

Korop M. M.

APPLICATION OF MACHINE LEARNING TO PREDICT THE PROPERTIES OF Bi_2Te_3-BASED THERMOELECTRIC MATERIALS

The paper provides examples of assessing the effectiveness of machine learning for predicting the properties of Bi_2Te_3-based thermoelectric materials. The results of their application and methods for selecting optimal input data parameters are considered, the differences and features of choosing algorithms, the stages of work and training machine models, as well as the criteria for assessing the effectiveness and validation of the obtained forecasts are described. Bibl. 18, Fig. 1, Tabl. 3.
Key words: machine learning methods, thermoelectric materials science.

Introduction

General characterization of the problem. Thermoelectric materials, which serve as the basis for machine-free thermal energy into electrical energy conversion, are gaining more and more popularity and practical application. One of the main criteria for evaluating promising materials is the quality factor Z proposed by Joffe, which can be expressed using formula 1:

$$ZT = \alpha^2 \frac{\sigma T}{\chi}, \quad (1)$$

where α is the Seebeck coefficient, σ is electrical conductivity, T is temperature, χ is thermal conductivity.

One of the most popular thermoelectric materials is Bi_2Te_3, the high figure of merit of which was achieved based on physical observations and an empirical approach. The increase in Z for this material occurred through the use of isovalent impurities, owing to which a decrease in χ (thermal conductivity) is achieved without a significant change in σ (electrical conductivity). During the following decades, further significant increase in the figure of merit did not occur, so the search for new methods and approaches is important.

In materials science, machine methods of optimizing materials and achieving their extreme values are of increasing interest. Machine learning is considered a subspecies of artificial intelligence that allows generalization, interpolation and extrapolation of input data, search for patterns and operating information in a more intelligent way. The main task of machine learning in predicting the properties of thermoelectric materials is to find the most accurate values with the smallest error, based on a limited amount of input data, which serve as a source of information obtained using both theoretical calculations and experimental measurements. Therefore, the question was raised to study how effective machine learning is for its application in thermoelectric materials science and to consider a number of works on its application.

The purpose of the work is to study the efficiency of machine learning methods in the task of predicting and optimizing the properties of Bi_2Te_3-based thermoelectric materials.

Supervised machine learning methods for Bi_2Te_3

Supervised machine learning methods are one of the main types of machine learning, where "supervised" means that the learning process is based on data with labels (absolute values) of the desired variable. Such methods are often used to solve problems of regression (predicting a numerical value) or classification (assigning data to their category). The accuracy and efficiency of such methods heavily depends on the quality of the training data and avoiding the risk of overtraining the model (having adapted to the input data, the model stops generalizing the information received and making accurate predictions for new data that the model has not previously received).

The following controlled machine learning algorithms have been used in scientific works devoted to solving problems of materials science: linear regression, logistic regression, decision trees, random forest, support vector machines, neural networks. Each of the algorithms has its own advantages and areas of application, and one of the main factors of choosing among them is the achievement of the highest accuracy under the given conditions [2].

Regression algorithms of machine learning

In the field of machine learning, regression algorithms are statistical methods that make a numerical prediction for a dependent variable based on its dependence on one or more independent variables.

For example, when predicting the *ZT* (thermoelectric figure of merit) value, it (the *ZT* value) will act as a dependent variable, and the input data of temperature, electrical conductivity, and Seebeck coefficient will act as independent variables.

There are a number of approaches to solving regression problems: linear regression, logistic regression, polynomial, Ridge regression, and Lasso regression [15]. Such models are relatively easy to learn, do not require large computing power, but contain limitations in the ability to generalize only simple dependences.

Linear regression models detect a linear relationship between *x* and *y* based on an input statistical data set. The mathematical representation of such a model can be described using formula 2.

$$y = b + \sum_{i=1}^{n} w_i * x_i, \tag{2}$$

where *y* is the predicted value of the dependent variable;
b (bias) is an absolute value that allows the model to account for bias in the output values that cannot

be explained by the independent variables; *w* is the weighting factor, which indicates how much the change in the independent variable *x* explains the change in the dependent variable *y*.

Polynomial regression models are designed to find non-linear relationships between input and output values. Such models are able to generalize more complex cases and are described using formula 3.

$$y = b + \sum_{i=0}^{n} w_i * x_i^i, \tag{3}$$

The risk of using polynomial regression is that the model will overtrain with high values of *n*, high prediction accuracy on training data, and low accuracy with new data and difficulty in explaining the trained model.

One of the main reasons for overtraining the model is too large values of the coefficients in equations 2 and 3 for the input parameters (weights) and a small amount of training data. To solve the problem of insufficient training data, data expansion methods can be used, namely: introducing noise for numerical values, changing the size, brightness for images, etc. Here, the term "noise" should be understood as random errors or variations to the original data - they allow providing training data to the input of the model that are as close as possible to real values. In the case of retraining the model, as a precautionary measure, regularization is used, which is a control method and adds additional constraints on the model weights. This includes L_1 (Lasso), L_2 (Ridge), or a combination of these regularizations.

At each stage of model training, cost functions are applied to achieve maximum accuracy. Their main task is to estimate the model error during training to adjust the coefficients (weights) so that this error is minimal. For regression problems, the root mean square error is often used, which is described by formula 4. Based on the values of the cost function, gradients are calculated to calculate new weights for each training step.

$$\sigma = \frac{\sum_{i=1}^{n}(x_i - y_i)^2}{n}, \tag{4}$$

where *x* is the exact value, *y* is the value predicted by the model.

Thus, the L_1 and L_2 regularizations mentioned above act as additional terms that are added to the cost function during model training. L_1 regularization adds absolute values of weights for certain descriptors, thereby changing them to 0 for parameters that are insignificant. L_2 regularization adds squared weights to the cost function, thereby "penalizing" the model for large weight values.

A practical way to apply supervised machine learning methods

The organization of the learning process of an artificial intelligence model can be represented as the following list of actions.

1. **Collection of training data:** at this stage, primary data (independent variables) are collected, which can be expressed by numerical values (crystal structure, chemical formula, average atomic number, etc.) or categorical values [3]. Dependent values (labels) that represent the outcome or class to be predicted are obtained from theoretical calculations or experimental measurements.
2. **Pre-processing of data:** aims to improve the quality of collected data, find and remove noise, missing or incorrect data. Categorization and coding of data or their normalization is carried out [4].
3. **Selection of the model algorithm:** is carried out on the basis of prepared data in such a way as to achieve the set goals and obtain the highest accuracy. The selection process is highly dependent on the task at hand and the input data set. Evaluation of input data, presence of noise, number of

functions, degree of linearity between variables, trade-off between complexity and efficiency of the model, minimizing the possibility of retraining the model.

4. **Training of the selected model and optimization**: adjustments (weights, thresholds, etc.) to achieve accuracy targets and minimize errors.
5. **Evaluation of the performance of trained model:** the goal is to identify and correct problems that arose during the training process, for example, accurate work with training data and incorrect work with new sets of input data.
6. **Testing of the resulting model**: carried out to determine its accuracy and efficiency.

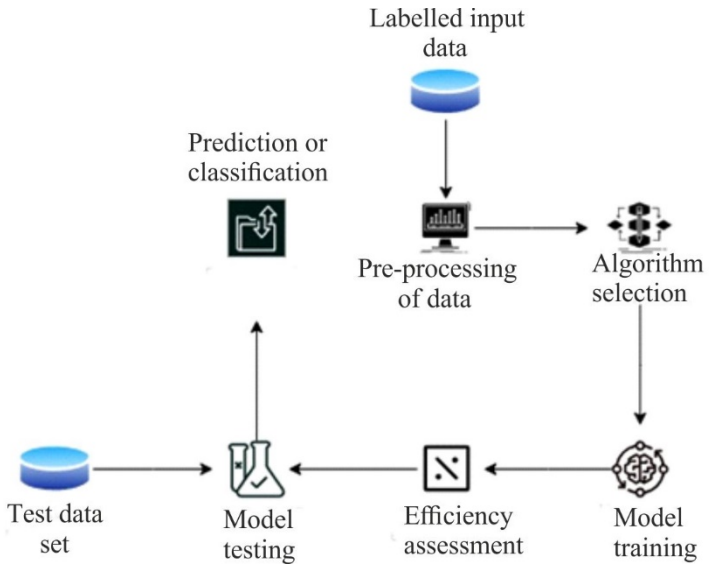

Fig. 1. Scheme of application of a supervised machine learning method

Wudil, Y. et al. [1] carried out work on predicting the quality factor of Bi_2Te_3-based materials using the lattice structure constants (a and c) and the electrical properties of the materials as predictors. The input data was generated based on a number of experimental works, from which the transport properties and crystal lattice parameters were taken. Also, to generalize the model, the results were formed both for the pure material and for the material with impurities. Two intensively researched compounds of *n*-type $Bi_2Te_{2.7}Se_{0.3}$ and *p*-type $Bi_{1.5}Sb_{0.5}Te_3$ were taken into account. The final data set contained 280 data points.

5 parameters were selected as descriptors: Seebeck coefficient, electrical conductivity, temperature, lattice dimensions (a, c). These parameters were selected by finding the interdependence of the target value and the input data, determined using the Pearson correlation coefficient. The choice of lattice parameters is expedient due to the dependence of the structural parameters of the lattice on the methods of manufacturing the material.

Parv Katyal, et al.[5] in their study set themselves the goal of determining algorithms and methods of machine learning that can work with a limited set of input data to establish interdependences between the *ZT* coefficient and the chemical and physical properties of compounds Bi_2Te_3, $CoSb_3$, $Ba_8Ga_{16}Ge_{30}$. To train the artificial intelligence, a data set of 1098 calculated points was used, in this set the *ZT* values were obtained at different temperatures. The descriptors chosen were: temperature, Seebeck coefficient, power factor, cell volume, resistance, total mass of one cell, average atomic volume, space group,

symmetry elements. The data set was divided randomly in the proportion of 80 % training data and 20 % testing data. For machine learning, the random forest method was selected, which is an ensemble method that consists of constructing several decision trees during training and as a result of calculating the average prediction of individual trees [6]. The authors of [5] divided the data set into 500 decision trees and calculated the average value of the results of all trees.

Zhi-Lei Wang, et al. [7] conducted a study to predict the properties of extruded $Cu_xBi_2Te_{2.85}Se_{0.15}$ samples using machine learning. A data set of seven experimental data sets and 12 characteristics were used: composition, relative density, orientation factor, average grain size, microstructural feature, grain boundary characteristics, charge carrier concentration, mobility, Seebeck coefficient, electrical resistance, thermal conductivity. An artificial neural network with one hidden layer was selected as the modeling algorithm, the sigmoid function was applied as the activation function, and the hyperparameters hidden layer node size and weight reduction were optimized using Bayesian optimization at a learning rate of 0.01. To avoid overtraining of the model, the input data was divided in the proportion of 60 % - training and 40 % - test using 4-fold cross-validation [8]. Cross-validation helps reduce the risk of model overtraining by dividing the data set into several equal or nearly equal parts. After this, the model is trained on all parts of the data, and the last part is left for testing. This process is repeated several times, each time selecting a different part as the test set. As a result, average indicators are calculated and an overall performance rating is found.

Qu, R. [16] investigates the thermal conductivity properties of $MnBi_2Te_4$ and $Bi_2Te_3/MnBi_2Te_4$ superlattice by applying a deep neural network (DNNP) using data obtained using density functional theory. Using DFT calculations, datasets containing atomic configurations, corresponding energies and forces, and temperatures (ranging from 200 K to 500 K) were prepared for model training. For training reliability, 1200 configurations were prepared, divided evenly across the specified temperature range, providing a comprehensive representation of the class space of the system. The DNNP model was structured with an embedded layer followed by three hidden layers, each with 160 nodes. This architecture was chosen to effectively capture the complexity of atomic interactions in the materials being studied. The training process was carefully configured, optimizing loss functions and hyperparameters to minimize the error between DNNP predictions and DFT calculations.

Research by Agarwal A. et al. [17] discusses a method for predicting the thermoelectric power factor for Bi_2Te_3 material during a powder bed laser melting (PBF-LB) process using machine learning techniques for additive manufacturing (AM). Additive manufacturing (AM) refers to a group of technologies that create objects by adding material layer by layer, based on digital 3D models. Unlike traditional subtractive manufacturing methods, which start with a solid stock and then cut away the excess to create the part, AM builds parts directly from the raw material layer by layer, which minimizes waste and allows the creation of complex geometries that would be difficult or impossible to achieve with conventional production methods. Ensemble machine learning methods are used to predict the power factor of Bi_2Te_3. Using specialized equipment, processing parameters and sensor data such as laser power, speed, layout spacing, layer thickness, and focus were collected. For image processing, the OpenCV Python library was applied to transform the acquired sensor images into meaningful features, including texture, surface roughness, and pixel intensity statistics. Feature scaling and data partitioning (80 % for training, 20 % for testing) were performed to ensure model reliability and generalization.

Headley, C.V. et al. [18] in his work describes an innovative combination of machine learning (ML) techniques with additive manufacturing (AM) processes to optimize the production of thermoelectric materials. In particular, the work draws attention to the production of parts from *n*-type

$Bi_2Te_{2.7}Se_{0.3}$ using a laser for melting and fusion of powder material into solid 3D objects. The study uses support vector regression (SVR) within ensemble learning. A bootstrap method was used to build an ensemble of SVR models, a technique that involves resampling from the training data set with replacement to train multiple models. The main descriptors for these models are LPBF process parameters such as laser power (which varied from 10 W to 40 W) and scan speed (which varied from 250 mm/s to 550 mm/s), which directly affect the width and depth of the molten of the "pool" area. To evaluate the model, the mean and standard deviation of forecasts from the ensemble of SVR models were used. These metrics are important for understanding the accuracy and reliability of ML predictions for the size of the molten "pool" that forms during the LPBF process.

Unsupervised machine learning methods for Bi_2Te_3

Unsupervised machine learning methods are a separate category of algorithms capable of learning from input data without direct control and defined labels [9]. Such methods find patterns of behaviour in data and their relationships in an offline mode, which allows one to effectively solve the tasks of clustering or finding associations. Unsupervised learning is a universal method, as it can find non-obvious connections of data in complex structures. Popular algorithms include: κ-means, hierarchical clustering, principal component algorithm, neural networks. Table 1 presents a visual comparison of supervised and unsupervised machine learning methods.

Table 1

Comparison of supervised and unsupervised machine learning methods

Supervised machine learning methods	**Unsupervised machine learning methods**
Prediction of a numerical value or classification of input data with labels.	Finding patterns and relationships in data based on data without any associated labels.
It requires input data processing, normalization, possible label encoding, feature selection, or development of specific features for the prediction task.	The main emphasis is on the selection of features to determine the more fundamental qualities of the data; there is no implementation of labels and their processing.
The model is selected depending on the task (regression or clustering), key parameters of the model are adjusted to minimize the difference between the actual and predicted labels.	Based on the type of pattern recognition (association, clustering, dimensionality reduction), the desired algorithm is selected without reference to labelled input data.
Absolute and relative error, precision, root mean square error are used as parameters for evaluating the efficiency of learning models. Input data is divided into educational and training data.	Silhouette index and Davis-Boldin index for clustering problems, as well as a subjective assessment of the membership of each selected instance with other objects in the class.
The obtained results are interpreted in the context of the given task and possible re-calibration of the model based on the performance evaluation.	There is a need to visualize the obtained results for better interpretation and focus on understanding the identified groupings and patterns.

Summarizing the table above, supervised machine learning methods can be characterized as such that work with labelled input data and allow solving prediction and clustering problems. In turn, unsupervised machine learning methods guide the research process to identify patterns and hidden structures of interdependences in unlabelled input data.

At the moment, no works have been found that use unsupervised machine learning methods to study the Bi_2Te_3 thermoelectric material, but there are several interesting areas of their application:

1. Formulation of the clustering problem for grouping samples of Bi_2Te_3 materials according to their similarity to a certain class, which will allow to reveal regularities and correlations in large data sets regarding the influence of synthesis methods, addition of impurities or nanostructure on the performance of such samples. Such a study would help to develop a methodology and recommendations for improving the process of synthesis and processing of material to solve the problems [10].
2. Application of unsupervised learning algorithms to extract features from microscopic images or crystallographic data of Bi_2Te_3 to find and cluster defects or crystal structures. The results of the work can be used as a way to improve the quality of control over the production of thermoelectric material and assess the effect of defects on the efficiency of samples [11].
3. Collection and preparation of a data set of impurities that are used and traditionally unexplored with Bi_2Te_3 to search for promising combinations. Due to the high performance of machine learning and the specific use of unsupervised methods, it is possible to develop a significant number of possible compounds [12].

Results of application of machine learning Bi_2Te_3

In the article by Wudil et al. [1] there were developed five weak regression models (Lasso regression, linear regression, decision tree regression, support vector regression) and one strong model combining the previous five using an ensemble technique using AdaBoost [13]. To evaluate the performance of these models, correlation coefficients, mean absolute error, coefficient of determination R2 and root mean square error were used. As a result, decision tree regression and support vector regression models showed high correlation coefficients of 99 % and 90.8 %, respectively. Enhanced models, using the AdaBoost algorithm, showed even higher indicators of 99.5 % and 94 %. During the validation, it was emphasized that the decision tree regression and support vector regression models with reduced mean absolute error and root mean square error are effective in assessing material quality. On the basis of this study, it is concluded that the implementation of enhanced weak regression algorithms significantly improves the accuracy of forecasting Bi_2Te_3 based thermoelectric semiconductors.

Another paper by Wudil [14] presents a scientific study using machine learning to estimate the thermal conductivity of materials based on $Bi_2Te_{2.7}Se_{0.3}$. Decision tree regression and support vector regression algorithms boosted with adaptive AdaBoost boost were also used in this work. For the selection of descriptors, a correlation was found between the input parameters of the data set and the sought value, presented in Table 2.

Table 2

Correlation coefficients between the input parameters and the target variable

	σ (S/m)	S (μV/K)	a (A)	c (A)	K (W/mK)	T (K)
σ (S/m)	1	–0.71	–0.15	–0.57	0.62	–0.36

Continue of table 2

S (μV/K)	–0.71	1	–0.22	0.33	–0.74	–0.088
a (A)	–0.15	–0.22	1	0.4	–0.11	0.076
c (A)	–0.57	0.33	0.4	1	–0.31	0.25
K (W/mK)	0.62	–0.74	–0.11	–0.31	1	0.29
T (K)	–0.36	–0.088	0.076	0.25	0.29	1

The model uses electrical properties and structural parameters of material lattices as input characteristics. The efficiency of the developed models is evaluated on the basis of such parameters as the correlation coefficient, the average absolute error, and the root mean square error. The decision tree model with AdaBoost enhancement showed a correlation coefficient of 99.4 % and a coefficient of determination R2 of 98.8 % in the test phase. These models have also been used to predict thermal conductivity for various physical specimens, such as transition metal compounds. The influence of substrate temperature during pulsed laser deposition was studied.

In the work of Parv Katyal et al. [5] the results of the study confirmed the high efficiency of predicting the *ZT* value for various compounds (Bi_2Te_3, $CoSb_3$, $Ba_8Ga_{16}Ge_{30}$, $Ba_8Ga_{18}Ge_{28}$) using the decision tree random forest algorithm. The evaluation of the efficiency of the model showed a low discrepancy with the expected result and an average absolute error of 0.0734, which shows the promise of this method in the processes of thermoelectric material evaluation. Table 3 presents the results of the proposed model for predicting *ZT* at different temperatures for a group of compounds.

Table 3

Experimental and predicted ZT values at different temperatures for the lead telluride family, the cobalt antimonide family, and the germanium-based clathrates

Temperature (K)	Experimental *ZT* value	ZT value predicted by the model	Chemical formula of material
400	0.5025	0.5923415	Bi_2Te_3
700	1.392715388	1.4253434	Bi_2Te_3
1000	1.636067789	1.5812441	Bi_2Te_3
600	0.871875767	0.938096442	Bi_2Te_3
300	0.316428584	0.384607488	Bi_2Te_3
300	0.424000225	0.502502554	$CoSb_3$
400	0.668512792	0.578206119	$CoSb_3$
700	1.181566055	1.16347007	$CoSb_3$
700	0.668168	0.7181711	$Ba_8Ga_{16}Ge_{30}$
300	0.01603609	0.067494695	$Ba_8Ga_{18}Ge_{28}$
1000	0.962666667	1.1242445	$Ba_8Ga_{18}Ge_{28}$

In the work of Wang Z. et al. [7] the study showed that the addition of copper impurities to *Bi-Te-Se* materials improves their thermoelectric characteristics. Copper atoms are introduced into interstitial spaces, changing the microstructure of the material and reducing the concentration of charge carriers. This leads to an increase in the Seebeck coefficient, electrical resistance and a decrease in the thermal conductivity of the media. The paper uses machine learning methods, including artificial neural network (ANN) and Bayesian optimization models, to predict and optimize the thermoelectric properties of these materials. Although the machine learning model is promising, problems related to retraining due to the small sample size have been noted.

In the work of Qu, R. [16], the model achieved high accuracy, as evidenced by low rms error values of 0.15 meV per atom for superlattice configurations for energy and force predictions. Predictions of thermal conductivity for $MnBi_2Te_4$ coincided well with experimental values, confirming the high efficiency of the DNNP model. The small discrepancies that were noted were within acceptable limits, highlighting the difficulties in excluding lattice from electronic thermal conductivity in experimental measurements. The $Bi_2Te_3/MnBi_2Te_4$ superlattice showed significantly reduced transverse thermal conductivity, which may be important for potential thermoelectric applications. In particular, the transverse thermal conductivity at 300 K was predicted to be 0.15 W $m^{-1}K^{-1}$, significantly lower than pure $MnBi_2Te_4$ or Bi_2Te_3, demonstrating the higher thermoelectric potential of the superlattice. Further analysis of the results showed that the decrease in transverse thermal conductivity in the superlattice can be explained by the dispersion relations of phonons, in particular, the appearance of bandgaps and a decrease in the speed of phonons. These phonon behaviours are key to understanding the mechanisms governing heat transfer in these complex materials.

Agarwal, A. et al. [17]. The classifier model based on bagging aggregation showed a high accuracy of 90 %, indicating a significant correlation between the selected features and the power factor of the thermoelectric material. Laser focus, power, and speed were among the main processing parameters affecting the power factor. Features associated with polarimetry data, especially post-distribution and post-melting angle of polarization (AoP) and degree of linear polarization (DoLP), were critical for power factor prediction. A total of 220 samples were produced and 117 were used for analysis, resulting in 3.157 data points for building machine learning models.

Headley, C. V. et al. [18] initially used 13 scan lines, which later expanded to 93 parameter combinations after six rounds of training, demonstrating the efficiency of an iterative, data-driven approach for refining process parameters. machine method, revealed the LPBF process parameters that resulted in the production of $Bi_2Te_{2.7}Se_{0.3}$ parts with a density greater than 99% and no cracks, demonstrating the high precision and quality that can be achieved. One notable advancement was the ability to produce thermoelectric parts with atypical geometries, such as hollow rectangles and trapezoids, with a relative density of 98.6 % (±1 %) and increased thermoelectric efficiency. The above shapes are difficult to produce using traditional manufacturing methods, but can be achieved through LPBF due to precise control over process parameters.

Conclusions

1. The efficiency of machine learning methods for predicting the properties of thermoelectric material Bi_2Te_3 was assessed.
2. Supervised machine learning algorithms, namely AdaBoost boosted weak models, decision tree regression, and support vector regression, are well suited for ZT factor prediction and thermal conductivity estimation of Bi_2Te_3-based thermoelectric materials.

3. Estimation of the efficiency of using the ensemble method, a random forest of decision trees, showed a low divergence with the expected result and an average absolute error of 0.0734 for compounds (Bi_2Te_3, $CoSb_3$, $Ba_8Ga_{16}Ge_{30}$, $Ba_8Ga_{18}Ge_{28}$).
4. There is considerable complexity in the amount of existing experimentally measured information about thermoelectric materials, which forces researchers to work with a limited set of information, which in turn leads to a decrease in the accuracy of forecasts.
5. With the help of machine learning, it is possible to determine a number of parameters for the efficient production of thermoelectric parts using additive manufacturing methods.

References

1. Wudil, Y. & Gondal, M. A. (2022). Predicting the thermoelectric energy figure of merit of Bi_2Te_3-based semiconducting materials: A machine learning approach. *SSRN Electronic Journal. Elsevier BV.* https://doi.org/10.2139/ssrn.4215166
2. Burkart, N. & Huber, M. F. (2021). A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, 70, 245 – 317. AI Access Foundation. https://doi.org/10.1613/jair.1.12228
3. Gaultois, M. W., Oliynyk, A. O., Mar, A., Sparks, T. D., Mulholland, G. J. & Meredig, B. (2016). Perspective: Web-based machine learning models for real-time screening of thermoelectric materials properties. *APL Materials*, 4 (5). AIP Publishing. https://doi.org/10.1063/1.4952607
4. Gonzalez Zelaya, C. V. (2019). Towards explaining the effects of data preprocessing on machine learning. *IEEE 35th International Conference on Data Engineering (ICDE).* IEEE. https://doi.org/10.1109/icde.2019.00245
5. Parv Katyal, Madhav Rathi, Piyush Mehra and Amrish K. Panwar (2020). Evaluation of figure of merit of thermoelectric materials using machine learning. *International Journal of Advanced Science and Technology*, 29(11s), 2858-2863. Retrieved from http://sersc.org/journals/index.php /IJAST/article/view/23766
6. Liu, Y., Wang, Y., & Zhang, J. (2012). New machine learning algorithm: random forest. In *Information Computing and Applications* (pp. 246 – 252). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-34062-8_32
7. Wang, Z., Yokoyama, Y., Onda, T., Adachi, Y., & Chen, Z. (2019). Improved thermoelectric properties of hot-extruded Bi–Te–Se bulk materials with Cu doping and property predictions via machine learning. *Advanced Electronic Materials*, 5 (6). Wiley. https://doi.org/10.1002/aelm.201900079
8. A. Ramezan, C., A. Warner, T., & E. Maxwell, A. (2019). Evaluation of sampling and cross-validation tuning strategies for regional-scale machine learning classification. *Remote Sensing*, 11(2), 185. MDPI AG. https://doi.org/10.3390/rs11020185
9. Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., & Aljaaf, A. J. (2019). A systematic review on supervised and unsupervised machine learning algorithms for data science. *Unsupervised and Semi-Supervised Learning* (pp. 3 – 21). Springer International Publishing. https://doi.org/10.1007/978-3-030-22475-2_1
10. Na, G. S. (2023). Artificial intelligence for learning material synthesis processes of thermoelectric materials. *Chemistry of Materials*, 35(19), 8272 – 8280). American Chemical Society (ACS). https://doi.org/10.1021/acs.chemmater.3c01834
11. Sheng, Y., Deng, T., Qiu, P., Shi, X., Xi, J., Han, Y., & Yang, J. (2021). Accelerating the discovery

of *Cu – Sn – S* thermoelectric compounds via high-throughput synthesis, characterization, and machine learning-assisted image analysis. *Chemistry of Materials,* 33(17), 6918 – 6924. American Chemical Society (ACS). https://doi.org/10.1021/acs.chemmater.1c01856

12. Jia, X., Deng, Y., Bao, X., Yao, H., Li, S., Li, Z., Chen, C., Wang, X., Mao, J., Cao, F., Sui, J., Wu, J., Wang, C., Zhang, Q., & Liu, X. (2022). Unsupervised machine learning for discovery of promising half-Heusler thermoelectric materials. *Computational Materials*, 8(1). Springer Science and Business Media LLC. https://doi.org/10.1038/s41524-022-00723-9
13. CAO, Y., MIAO, Q.-G., LIU, J.-C., & GAO, L. (2013). Advance and prospects of AdaBoost algorithm. *Acta Automatica Sinica,* 39 (6), 745–758. Elsevier BV. https://doi.org/10.1016/s1874-1029(13)60052-x
14. Wudil, Y. S. (2023). Ensemble learning-based investigation of thermal conductivity of Bi_2Te_2.7Se0.3-based thermoelectric clean energy materials. *Results in Engineering,* 18, 101203. Elsevier BV. https://doi.org/10.1016/j.rineng.2023.101203
15. Wang, T., Zhang, C., Snoussi, H., & Zhang, G. (2019). Machine learning approaches for thermoelectric materials research. *Advanced Functional Materials,* 30(5). Wiley. https://doi.org/10.1002/adfm.201906041
16. Qu, R., Lv, Y., & Lu, Z. (2023). A deep neural network potential to study the thermal conductivity of MnBi2Te4 and Bi2Te3/MnBi2Te4 superlattice, *Journal of Electronic Materials*, 52(7), 4475 – 4483). Springer Science and Business Media LLC. https://doi.org/10.1007/s11664-023-10403-z
17. Agarwal, A., Banerjee, T., Gockel, J., LeBlanc, S., Walker, J., & Middendorf, J. (2023). *Predicting thermoelectric power factor of bismuth telluride during laser powder bed fusion additive manufacturing (Version 1).* arXiv. https://doi.org/10.48550/ARXIV.2303.15663
18. Headley, C. V., Herrera del Valle, R. J., Ma, J., Balachandran, P., Ponnambalam, V., LeBlanc, S., Kirsch, D., & Martin, J. B. (2024). The development of an augmented machine learning approach for the additive manufacturing of thermoelectric materials. *Journal of Manufacturing Processes*, 116, 165 – 175). Elsevier BV. https://doi.org/10.1016/j.jmapro.2024.02.045

Submitted: 12.04.2023

Анатичук Л. І., *акад. НАН України* [1,2]
Короп М. М. [1]

[1] Інститут термоелектрики НАН та МОН України,
вул. Науки, 1, Чернівці, 58029, Україна;
e-mail: anatych@gmail.com, mykola.korop@chnu.edu.ua
[2] Чернівецький національний університет
імені Юрія Федьковича,
вул. Коцюбинського 2, Чернівці, 58012, Україна
e-mail: anatych@gmail.com

ЗАСТОСУВАННЯ МАШИННОГО НАВЧАННЯ ДЛЯ ПРОГНОЗУВАННЯ ВЛАСТИВОСТЕЙ ТЕРМОЕЛЕКТРИЧНИХ МАТЕРІАЛІВ НА ОСНОВІ Bi_2Te_3

У роботі наводяться приклади оцінки ефективності застосування машинного навчання для прогнозування властивостей термоелектричних матеріалів на основі Bi_2Te_3. Оглянуто результати їх застосування та способи вибору оптимальних параметрів вхідних даних, описано відмінності та особливості вибору алгоритмів, етапи роботи та навчання машинних моделей, а також критерії оцінки ефективності та валідації отриманих прогнозів. Бібл. 18, рис. 1, табл. 3.

Ключові слова: методи машинного навчання, термоелектричне матеріалознавство.

References

1. Wudil, Y. & Gondal, M. A. (2022). Predicting the thermoelectric energy figure of merit of Bi_2Te_3-based semiconducting materials: A machine learning approach. *SSRN Electronic Journal. Elsevier BV.* https://doi.org/10.2139/ssrn.4215166
2. Burkart, N. & Huber, M. F. (2021). A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, 70, 245 – 317. AI Access Foundation. https://doi.org/10.1613/jair.1.12228
3. Gaultois, M. W., Oliynyk, A. O., Mar, A., Sparks, T. D., Mulholland, G. J. & Meredig, B. (2016). Perspective: Web-based machine learning models for real-time screening of thermoelectric materials properties. *APL Materials*, 4 (5). AIP Publishing. https://doi.org/10.1063/1.4952607
4. Gonzalez Zelaya, C. V. (2019). Towards explaining the effects of data preprocessing on machine learning. *IEEE 35th International Conference on Data Engineering (ICDE).* IEEE. https://doi.org/10.1109/icde.2019.00245
5. Parv Katyal, Madhav Rathi, Piyush Mehra and Amrish K. Panwar (2020). Evaluation of figure of merit of thermoelectric materials using machine learning. *International Journal of Advanced Science and Technology*, 29(11s), 2858-2863. Retrieved from http://sersc.org/journals/index.php /IJAST/article/view/23766
6. Liu, Y., Wang, Y., & Zhang, J. (2012). New machine learning algorithm: random forest. In *Information Computing and Applications* (pp. 246 – 252). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-34062-8_32
7. Wang, Z., Yokoyama, Y., Onda, T., Adachi, Y., & Chen, Z. (2019). Improved thermoelectric properties of hot-extruded Bi–Te–Se bulk materials with Cu doping and property predictions via machine learning. *Advanced Electronic Materials*, 5 (6). Wiley. https://doi.org/10.1002/aelm.201900079
8. A. Ramezan, C., A. Warner, T., & E. Maxwell, A. (2019). Evaluation of sampling and cross-validation tuning strategies for regional-scale machine learning classification. *Remote Sensing*, 11(2), 185. MDPI AG. https://doi.org/10.3390/rs11020185
9. Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., & Aljaaf, A. J. (2019). A systematic review on supervised and unsupervised machine learning algorithms for data science. *Unsupervised and Semi-Supervised Learning* (pp. 3 – 21). Springer International Publishing. https://doi.org/10.1007/978-3-030-22475-2_1
10. Na, G. S. (2023). Artificial intelligence for learning material synthesis processes of

thermoelectric materials. *Chemistry of Materials*, 35(19), 8272 – 8280). American Chemical Society (ACS). https://doi.org/10.1021/acs.chemmater.3c01834

11. Sheng, Y., Deng, T., Qiu, P., Shi, X., Xi, J., Han, Y., & Yang, J. (2021). Accelerating the discovery of *Cu – Sn – S* thermoelectric compounds via high-throughput synthesis, characterization, and machine learning-assisted image analysis. *Chemistry of Materials,* 33(17), 6918 – 6924. American Chemical Society (ACS). https://doi.org/10.1021/acs.chemmater.1c01856
12. Jia, X., Deng, Y., Bao, X., Yao, H., Li, S., Li, Z., Chen, C., Wang, X., Mao, J., Cao, F., Sui, J., Wu, J., Wang, C., Zhang, Q., & Liu, X. (2022). Unsupervised machine learning for discovery of promising half-Heusler thermoelectric materials. *Computational Materials*, 8(1). Springer Science and Business Media LLC. https://doi.org/10.1038/s41524-022-00723-9
13. CAO, Y., MIAO, Q.-G., LIU, J.-C., & GAO, L. (2013). Advance and prospects of AdaBoost algorithm. *Acta Automatica Sinica,* 39 (6), 745–758. Elsevier BV. https://doi.org/10.1016/s1874-1029(13)60052-x
14. Wudil, Y. S. (2023). Ensemble learning-based investigation of thermal conductivity of Bi_2Te_2.7Se0.3-based thermoelectric clean energy materials. *Results in Engineering,* 18, 101203. Elsevier BV. https://doi.org/10.1016/j.rineng.2023.101203
15. Wang, T., Zhang, C., Snoussi, H., & Zhang, G. (2019). Machine learning approaches for thermoelectric materials research. *Advanced Functional Materials,* 30(5). Wiley. https://doi.org/10.1002/adfm.201906041
16. Qu, R., Lv, Y., & Lu, Z. (2023). A deep neural network potential to study the thermal conductivity of MnBi2Te4 and Bi2Te3/MnBi2Te4 superlattice, *Journal of Electronic Materials*, 52(7), 4475 – 4483). Springer Science and Business Media LLC. https://doi.org/10.1007/s11664-023-10403-z
17. Agarwal, A., Banerjee, T., Gockel, J., LeBlanc, S., Walker, J., & Middendorf, J. (2023). *Predicting thermoelectric power factor of bismuth telluride during laser powder bed fusion additive manufacturing (Version 1).* arXiv. https://doi.org/10.48550/ARXIV.2303.15663
18. Headley, C. V., Herrera del Valle, R. J., Ma, J., Balachandran, P., Ponnambalam, V., LeBlanc, S., Kirsch, D., & Martin, J. B. (2024). The development of an augmented machine learning approach for the additive manufacturing of thermoelectric materials. *Journal of Manufacturing Processes*, 116, 165 – 175). Elsevier BV. https://doi.org/10.1016/j.jmapro.2024.02.045

Submitted: 12.04.2023