Journal of Applied Data Sciences
Vol. 5, No. 4, December 2024, pp. 1914-1924

An Improved Prediction of Transparent Conductor Formation Energy using PyCaret: An Open-Source Machine Learning Library

Ayorinde Tayo Olanipekun[1,*], Daniel Mashao[2]

[1]*Data Science Across Disciplines Research Group, Institute for the Future of Knowledge, Faculty of Engineering and the Built Environment, University of Johannesburg, Auckland Park, South Africa*

[2]*Faculty of Engineering and the Built Environment, University of Johannesburg, South Africa*

(Received: April 21, 2024; Revised: May 13, 2024; Accepted: July 11, 2024; Available online: November 7, 2024)

Abstract

Designing innovative materials is necessary to solve vital energy, health, environmental, social, and economic challenges. Transparent conductors are compounds that combine low absorption visible range and good electrical conductivity, which are essential properties for conductors. Technological devices such as photovoltaic cells, transistors, photovoltaic cells and sensors majorly rely on combining the two properties due to their relevancy in an optoelectronic application. Meanwhile, fewer compounds exhibit both outstanding conductivity and transparency suitable for their application in transparent conducting materials. Kaggle hosted an open big-data competition organized by novel material discovery (NOMAD) to address the importance of finding new material with the ideal functionality. The competition was organized to identify the best machine learning (ML) to predict formation enthalpy (indicating stability) for 3000 $(Al_xGa_yIn_z)_{2N}O_{3N}$ compounds datasets; where x, y, and z can vary from the constraints x+y+z=1. Here we present a prediction using an open-source machine learning library in Python called PyCaret to summarise top-ranked ML algorithms. The gradient boosting regressor (GBR) model performed best with MAE 0.0281, MSE 0.0018 and R^2 0.84. The research shows that Machine learning can significantly accelerate the discovery and optimization of materials while reducing cost of computation and required time. Low code tools like PyCaret were used to enhance the machine learning applications in materials science, paving way for more efficient materials discovery processes.

Keywords: Transparent Conductor, Machine Learning, PyCaret, Open Source, Data Science

1. Introduction

The development of innovative materials can be challenging whenever it concerns energy, health equipment, and many other fields. Therefore, for material properties to be optimized, it is essential to deeply understand the relationship among composition, properties, and internal energy conditions. More importantly, in our consideration, transparent conductors are significant compounds with low absorption in the visible range and are electrically conductive. These important and unique properties of transparent conductors make them applicable in sensors, laser equipment and transistors [1].

The foundation of any innovation in any industry is the development of functional materials. Advanced computational techniques, such as density functional theory (DFT), have been helpful to a certain extent in this regard. DFT requires many computing resources and time to achieve the set goal [2], [3], [4], [5], [6]. ML will be used to predict the essential features of the transparent conductor in this research as opposed to DFT. ML usually accelerate novel materials discovery by effectively screening candidate compounds using data analysis at impressively lower computational cost when compared to traditional electronic structure approaches [7], [8], [9], [10], [11], [12]. Meanwhile, ML has passed as an important tool in computational chemistry to fast-track materials design and atomistic simulations. ML has the probability to enhance the computational efficiency of electronic structure and its predictive capacity has corrected any errors that could arise from the density functional approach [13].

Big data-driven science has been said to belong to the "fourth paradigm" for scientific exploration, while experiments, theory and simulation make the first three on the chart [14]. As the value of data-intensive approaches becomes

***Corresponding author: Ayorinde Tayo Olanipekun (atolanipekun@uj.ac.za)**
DOI: https://doi.org/10.47738/jads.v5i4.202

imminent, the materials science field has not yet witnessed much adoption of these methods as compared to other fields like astronomy [15], biosciences [16] and particle science [17]. Meanwhile, the advantages of material-driven science cannot be overemphasized: Material development cycle and commercialization that can take close to 10 until 20 years can be effectively reduced by material informatics [17], [18].

Hence, in this research, PyCaret, an open-source, low-code machine learning library in Python, was used to make a machine learning comparative analysis using standardized error metrics like MSE, R^2, and MAE. PyCaret is a straightforward and effective AutoML that boosts effectiveness and expedites ML research [19]. Seventeen ML model analyses are carried out on transparent conductors, containing datasets of 3000 conductors, with 12 features that were used to predict the critical properties of the conductors, which is formation energy (Ef).

2. Literature review

Many studies have considered the prediction of the formation energy of transparent conductors using ML. This chapter will review some of the existing literature on the application of ML to the prediction of formation energy in transparent conductors. This chapter also reviews various ML models and techniques that have been employed to predict formation energy, highlighting key studies, methodologies and their outcomes. This review aims to contextualize the current state of research, identify gaps in the existing literature and pride a foundation for further research.

Chenebuah et al. [20] used different ML models to predict the Ef and bandgap energy (Eg) of perovskites, usually computed from DFT simulations. The research found that support vector regression (SVR) model performs the best in predicting the formation energy. They also noticed that the GBR and random forest regression (RFR) models give better energy bandgap prediction than the SVR model.

Faber et al. [21] were able to predict the formation energy for Elpasolite materials, where 90 out of 212 new structures were predicted to be on the convex hull. Also, Pilania et al. developed a machine-learning model to estimate over 1200 binary wurtzite superlattices' various physical parameters, including bandgap, elastic constants, and formation energy [22].

In the work of Mao et al. [23] ML was used to predict the formation energies of materials without detailed crystal structure information by identifying key features. A model was created that accurately predicts formation energies for many binary compounds, aligning well with experimental data. The model uses important atomic properties like electronegativity and bond energy, making it useful for predicting and classifying binary compounds on a large scale without needing crystal structure details.

3. Methodology

3.1. Data and Variables Influencing the Predicted Properties

For ML to be effectively applied in a material science problem, the quality of the datasets for materials properties prediction must be of a high standard. In this research, we source the data from NOMAD data 2018 Kaggle competition for Predicting the key properties of novel transparent conducting oxides [21], [24]. The case study focused on high-quality data provided for 3000 $(Al_xGa_yln_z)_{2N}O_{3N}$ compounds are materials that show promise as transparent conductors. 2,400 of the materials were used for the training set, with the remaining 600 used for the test set. Meanwhile, each material has 11 features that contribute to the predicted properties' outcome.

Variables influencing the formation energy and bandgap energy (targeted properties to be predicted) of the transparent conductors are described below:

Space group: It shows the category of transparent conductors by a label identifying the symmetry of the material

Relative Compositions: The basic structure and properties of the materials are affected by Al, Ga, and Ln

Lattice angles (α, β, γ): It shows the basic structural and material unit.

Number of the atoms (Al, Ga, In and O) in the unit cell: This usually influences the basic structure and properties of the transparent conductors.

Coordinate Information (x, y, z): The general material structure is shown by the atom in each sample.

Lattice Vectors $lv1, lv2, lv3$: It connects two lattice points in the material unit.

3.2. Formulation of the Governing Equation.

To formulate the predicted electronic properties (formation energy Ef), we can use the equation below to describe the scenario:

$$E = f(Z) \tag{1}$$

E will stand for Ef, and Z will stand for other input features described in section 2.

A low-code machine learning open-source program called PyCaret was utilized to do feature selection, data splitting, model selection, and hyperparameter tuning [19]. Python libraries like matplotlib, Seaborn, and Sci-kit learn are used for modeling and visualization. Formation energy is selected as the targeted feature while variables above were used as the input features. The detailed process of the machine learning analysis is shown in figure 1.

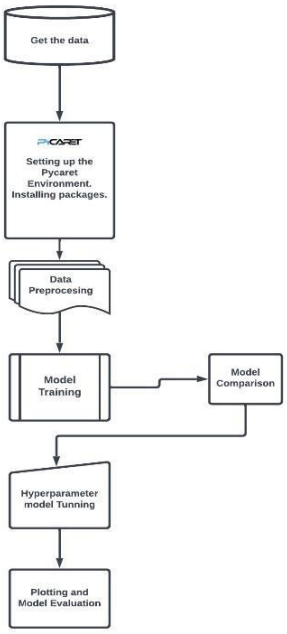

Figure 1. Flow chart for the Machine Learning process

3.3. Data Preprocessing

Data Preprocessing is very important to ensure input features are properly formatted. In this research, we tried to check for the possibility of missing values. Handling missing values is a very useful step in building reliable models. Proper data preprocessing improves model performance and reliability. The process data are then fed for the machine learning analysis. Overall, proper data preprocessing plays a pivotal role in improving the model's performance and reliability. By addressing potential data issues at this stage, we ensured that the processed data fed into the machine learning models was clean, consistent, and representative of the underlying patterns within the dataset. This rigorous preprocessing pipeline not only enhanced the quality of the data but also laid a solid foundation for subsequent machine learning analysis, leading to more accurate and dependable predictions.

3.4. Model Evaluation

3.4.1. Statistical Evaluation of the Model

Statistical analysis plays a vital role in evaluating and validating machine learning models, particularly in regression analysis where accurate prediction of continuous outcomes is crucial. R^2, MSE, and MAE statistical metrics are mostly used in regression analysis, for a trustworthy and comprehensible framework for assessing machine learning models, assisting practitioners in making defensible choices regarding model selection, optimization and implementation. The following statistical metrics were used to evaluate the performance of the algorithms in this work.

By using this comprehensive set of statistical metrics, we were able to rigorously evaluate the performance of our regression models, providing clear insights into their strengths and weaknesses. This analytical approach not only facilitated the selection of the most appropriate model but also guided the optimization process to achieve the best possible outcomes. Ultimately, these metrics and analyses form the backbone of our model evaluation framework, ensuring that the chosen models are both reliable and interpretable for practical implementation.

3.4.2. Coefficient of Determination (R^2)

Regression analysis uses the R-squared (R^2) statistic, sometimes referred to as the coefficient of determination, as a metric to evaluate how well a model fits the observed data. R^2 shows how effectively the model's independent variables account for variations in the dependent variable. R^2 usually ranges from 0 to 1. If the predicted outcome for R^2 is 1, it means the model is perfectly fitted to the data [25], [26]. The equation represents the formulation for:

$$R^2 = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i - \underline{y})^2} \quad (2)$$

In the above equation, y_i represents the actual value, $\underline{y}$ stands for the mean value, while y_i stands for the predicted value, i represents the index of each value and n is the number of observable features.

3.4.3. Mean Square Error (MSE)

The mean squared error (MSE) is the most widely used representation of the regression loss function. The loss is computed as the squared difference between the actual and predicted values averaged over all data points [27]. In practical terms, MSE is not only a measure of model performance but also serves as a diagnostic tool. By analyzing the MSE during model training and validation, practitioners can gain insights into how well the model is learning from the data and identify potential issues such as overfitting, underfitting, or the presence of noisy data. Additionally, comparing MSE values across different models or configurations provides a straightforward method for selecting the model that best balances accuracy and complexity. The Equation [28] for MSE is shown below:

$$\mathrm{MSE} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{y_i} - \hat{\mathrm{y}}_\mathrm{i})^2 \quad (3)$$

3.4.4. Mean Average Error (MAE)

It is used in regression problems to quantify the difference between the actual values and the predicted values. The smaller the MAE value the more accurate the prediction MAE provides a direct interpretation of the average error magnitude in the same units as the target variable [29]. The smaller the MAE, the closer the model's predictions are to the actual outcomes, indicating better accuracy.

The equation that represents MSE is shown below:

$$\mathrm{MSE} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}|\mathrm{y_i} - \hat{\mathrm{y}}_\mathrm{i}| \quad (4)$$

4. Results and Discussion

4.1. Data Visualization

The 3D plot is created in PyCaret to show outliers in the data. The 3D plot in figure 2, shows the outliers as those that do not belong to any cluster. The data sets are divided into 4 different clusters, so anything that deviates from those groupings will result in an anomaly. The dataset was divided into four distinct clusters. Clustering is a technique that groups data points based on their similarity, where points within the same cluster are more similar to each other than to those in different clusters. This grouping helps in understanding the natural structure of the data.

By visualizing the data in three dimensions, we can observe how the clusters are formed and how the data points are distributed across these clusters. Data points that lie outside these well-defined clusters are considered outliers, as they do not conform to the underlying distribution or pattern of the rest of the data. In the 3D plot, these outliers are easily distinguishable as they are positioned away from the dense regions representing the clusters.

Clusters: The data points are color-coded to represent the four different clusters. Each cluster is visually separated from the others, providing a clear distinction between the groups.

Outliers: Data points that do not belong to any of the identified clusters are marked as outliers. These outliers are significant because they deviate from the typical patterns observed within the clusters.

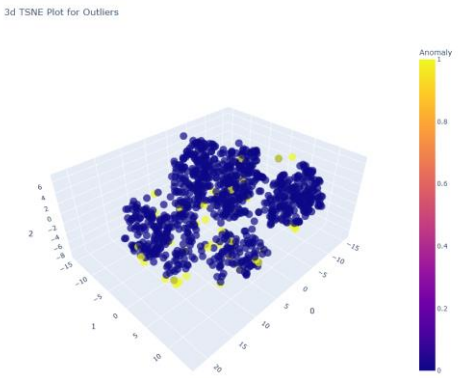

Figure 2. 3D TSNE Plot for Outliers

4.2. Importance of Different Independent Variables or Features

Feature importance analysis was calculated using the Gradient boosting classification model. From figure 3, Lattice Vectors ($lv3$) appear to be the most important feature in predicting the formation energy. Whereas, lattice angle has the least importance.

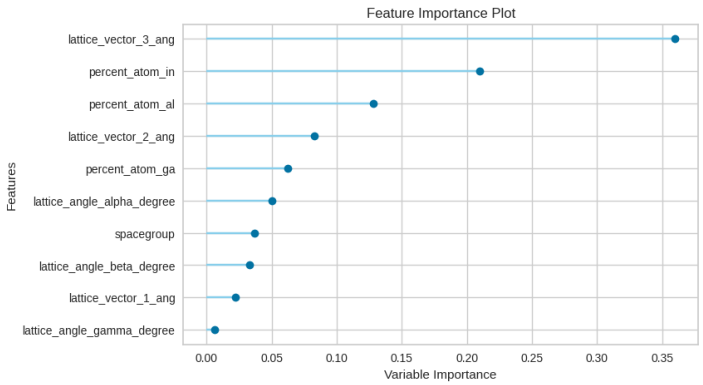

Figure 3. Feature Importance plot

The higher-ranking importance of the deviation weight compared to other features shows the importance of the lattice vector. Lattice vectors give basic details about the geometry and atomic organization of a crystal structure by defining

its unit cell. A material's atomic configuration, which is connected to the lattice vectors directly, determines its formation energy [30]

4.3. Machine Learning Analysis

Different machine learning algorithms were used for the formation energy prediction. The best-performed model was eventually chosen as shown in table 1. GBR and lightgbm show outstanding predictive capabilities. Figure 4 shows the regression plot, that shows the deviation of the predicted values when compared with the actual values. The statistical metrics R^2 value used to measure the performance of the model on training and test data was given as R^2=0.893for training and R^2=0.835 for testing.

4.3.1. Residual Plot

To further evaluate the performance of the GBR model, a residual plot was used. The plot as shown in figure 4, shows a residual plot of the predicted values (residuals) on the vertical axis against the predicted values on the horizontal axis.

From the plot, it can be noticed that the residuals are centered around zero, and only a few points deviate from the zero point, indicating a little bias in the model [31]. The randomness of the residuals is also one of the good qualities observed in the plot. The residual is randomly distributed around the horizontal axis of the plot [32].

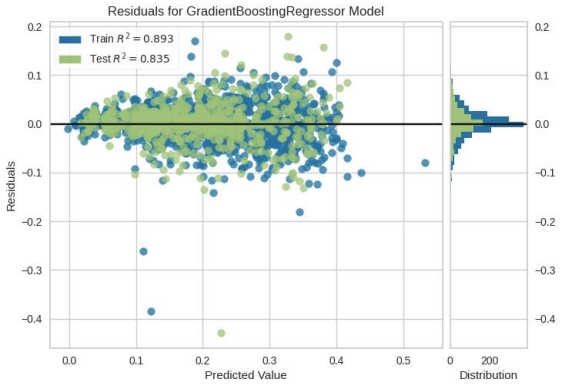

Figure 4. Residuals for GBR model

The randomness of the residuals is another positive attribute observed in the plot. Randomly distributed residuals around the horizontal axis imply that the model has captured the underlying patterns in the data without overfitting or underfitting. This randomness indicates that the errors are not correlated with the predicted values, which is essential for ensuring that the model generalizes well to unseen data.

Overall, the residual plot serves as a critical component in the model evaluation process, offering a visual and statistical validation of the GBR model's performance. The centralization of residuals around zero, coupled with their random distribution, highlights the model's effectiveness in making unbiased predictions, thus confirming its robustness and reliability for the task at hand. This analysis supports the continued use and potential optimization of the GBR model as a powerful tool in regression tasks within this research.

4.3.2. Prediction Error

The plot below represented by figure 5 shows the prediction error for GBR. R^2 value for the training is 0.835, this tells us that the error in our prediction is minimal. Assessing the prediction performance using the line of best fit, it can be observed that the model's predictions align somewhat with the true values, indicating a moderate level of accuracy. The identity line, conversely, illustrates that the data points are situated nearer to it, suggesting minimal error potential. This also indicates that our model's predictions closely match the true values. Furthermore, the line of best fit and the identity line showed minimal deviation from each other, suggesting little bias in the model's predictions [33].

 ISSN 2723-6471

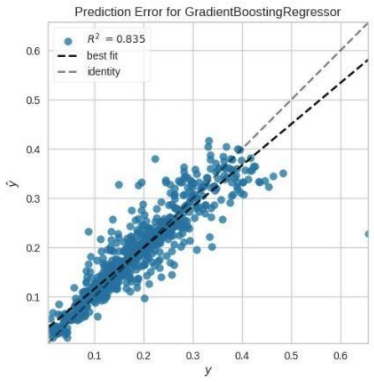

Figure 5. Prediction error for GBR model

4.3.3. 10-Fold cross-validation

For the cross-validation, the original datasets were randomly divided into 10 equals to size folds this is shown in table 1. The model was trained and evaluated 10 times for better prediction. This process is intended to reduce the variance in our prediction. As presented in table 1, the R^2 value 0.84 shows a little improvement in the prediction accuracy compared to the initial GBR model.

Table 1. GBR Performance for 10-fold cross-validation

Fold	MAE	MSE	R2
0	0.0308	0.0022	0.7991
1	0.0288	0.0018	0.8419
2	0.0253	0.0013	0.8760
3	0.0287	0.0022	0.8293
4	0.0266	0.0014	0.8856
5	0.0293	0.0017	0.8481
6	0.0308	0.0031	0.7204
7	0.0276	0.0014	0.8705
8	0.0269	0.0014	0.8569
9	0.0260	0.0014	0.8718
Mean	**0.0281**	**0.0018**	**0.8400**
Std	**0.0018**	**0.0005**	**0.0467**

The research has shown improvement in the prediction of some existing research on transparent conductor formation energy. The GBR predictions demonstrate enhanced accuracy, reaching 89%, surpassing the findings of Varadarajan et al. [34] and Christopher et al. [35]. Notably, our GBR model achieved a lower mean absolute error (MAE) value of 0.0281 than their reported MAE values as shown in table 2 [34], [35].

Table 2. Statistical Metrics for the different machine learning models

Model	MAE	MSE	R2
Gradient Boosting Regressor	0.0281	0.0018	0.8400
Light Gradient Boosting Machine	0.0281	0.0019	0.8341
Random Forest Regressor	0.0285	0.002	0.8246
Extra Trees Regressor	0.0291	0.0021	0.8125
Extreme Gradient Boosting	0.0303	0.0022	0.8021
Decision Tree Regressor	0.0352	0.0032	0.7141
AdaBoost Regressor	0.0477	0.0035	0.6836

Linear Regression	0.0638	0.0068	0.3958
Least Angle Regression	0.0638	0.0068	0.3958
Ridge Regression	0.0639	0.0068	0.3918
Bayesian Ridge	0.0639	0.0068	0.3916
K Neighbors Regressor	0.0742	0.0085	0.2437
Huber Regressor	0.0739	0.0088	0.2248
Orthogonal Matching Pursuit	0.0828	0.0100	0.1113
Elastic Net	0.0828	0.0100	0.1082
Lasso Regression	0.0832	0.0101	0.0994
Lasso Least Angle Regression	0.0832	0.0101	0.0994

5. Conclusion

In this research, PyCaret Machine learning low code Python package was used in the analysis. Different machine learning algorithms were used to predict the formation energy of the transparent conductor. GBR and lightgbm perform best. This approach can be used to design and screen many more transparent conductors, exploiting the formation energy properties. Precise formation energy prediction can stimulate the synthesis of transparent conductors.

Machine learning models are a significant development in the quickly evolving field of materials informatics that will help identify the next generation of energy materials. Meanwhile, our findings imply that ML models have enormous potential for the unprecedented rate and scope of computational screening of transparent conductors. The significance of using PyCaret for transparent conductor property prediction cannot be underemphasized, as it significantly reduces the amount of time it would take to experiment with different ML models, saving us precious time and ensuring quick iteration.

In future research, we can discard the features with low importance to effectively and accurately predict the formation energy. Removal of the redundant features will help us to potentially reduce computational cost and increase model accuracy. Also, in subsequent research, we plan to build a front-end app to host the machine learning API as a Web app for formation energy prediction.

6. Declarations

6.1. Author Contributions

Conceptualization: A.T.O. and D.M.; Methodology: D.M.; Software: A.T.O.; Validation: A.T.O., D.M.; Formal Analysis: A.T.O., D.M.; Investigation: A.T.O.; Resources: D.M.; Data Curation: D.M.; Writing Original Draft Preparation: A.T.O. and D.M.; Writing Review and Editing: D.M. and A.T.O.; Visualization: A.T.O. All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] K. Schütt, H. Glawe, F. Brockherde, A. Sanna, K. M. R. Müller, and R. Ballarin, "How to represent crystal structures for machine learning: Towards fast prediction of electronic properties," *Phys. Rev. B*, vol. 89, no. 20, pp. 205118-205129, Mar. 2014.

[2] J. Dean, M. Scheffler, T. A. R. Purcell, S. V. Barabash, R. Bhowmik, and T. Bazhirov, "Interpretable machine learning for materials design," *J. Mater. Res.*, vol. 38, no. 20, pp. 4477–4496, Oct. 2023, doi: 10.1557/s43578-023-01164-w.

[3] S. Curtarolo, G. L. W. Hart, M. B. Nardelli, N. Mingo, S. Sanvito, and O. Levy, "The high-throughput highway to computational materials design," *Nat. Mater.*, vol. 12, no. 3, pp. 191–201, Feb. 2013, doi: 10.1038/nmat3568.

[4] S. Chakraborty, M. L. Agiorgousis, M. H. Chen, and S. Huang, "Rational design: A high-throughput computational screening and experimental validation methodology for lead-free and emergent hybrid perovskites," *ACS Energy Lett.*, vol. 2, no. 4, pp. 837–845, Apr. 2017, doi: 10.1021/acsenergylett.7b00035.

[5] N. Mounet, M. Gibertini, P. Schwaller, D. Campi, A. Merkys, A. Marrazzo, T. Sohier, I. E. Castelli, and N. Marzari, "Two-dimensional materials from high-throughput computational exfoliation of experimentally known compounds," *Nat. Nanotechnol.*, vol. 13, no. 3, pp. 246–252, Mar. 2018, doi: 10.1038/s41565-017-0035-5.

[6] K. Kuhar, A. Crovetto, D. A. Saldana-Greco, J. M. Vela, R. Ramírez, and K. A. Persson, "Sulfide perovskites for solar energy conversion applications: Computational screening and synthesis of the selected compound $LaYS_3$," *Energy Environ. Sci.*, vol. 10, no. 12, pp. 2579–2593, Dec. 2017, doi: 10.1039/c7ee02702h.

[7] B. Meredig, A. Agrawal, S. Kirklin, J. E. Saal, J. W. Doak, A. Thompson, K. A. Wolverton, and C. Wolverton, "Combinatorial screening for new materials in unconstrained composition space with machine learning," *Phys. Rev. B*, vol. 89, no. 9, pp. 094104-094116, Mar. 2014, doi: 10.1103/PhysRevB.89.094104.

[8] O. Isayev, D. Fourches, E. N. Muratov, C. Tropsha, R. A. Cohn, D. R. Talley, C. S. Miller, M. Farhadifar, and S. Curtarolo, "Materials cartography: Representing and mining materials space using structural and electronic fingerprints," *Chem. Mater.*, vol. 27, no. 3, pp. 735–743, Feb. 2015, doi: 10.1021/cm503507h.

[9] A. O. Oliynyk, L. Antono, T. D. Sparks, L. Ghadbeigi, C. D. Gaultois, B. Meredig, and A. Mar, "High-throughput machine-learning-driven synthesis of full-Heusler compounds," *Chem. Mater.*, vol. 28, no. 20, pp. 7324–7331, Oct. 2016, doi: 10.1021/acs.chemmater.6b02724.

[10] J. Schmidt, J. Shi, P. Borlido, L. Chen, S. Botti, and M. A. L. Marques, "Predicting the thermodynamic stability of solids combining density functional theory and machine learning," *Chem. Mater.*, vol. 29, no. 12, pp. 5090–5103, Jun. 2017, doi: 10.1021/acs.chemmater.7b00156.

[11] A. Mannodi-Kanakkithodi, G. Pilania, T. D. Huan, T. Lookman, and R. Ramprasad, "Machine learning strategy for accelerated design of polymer dielectrics," *Sci. Rep.*, vol. 6, no. 1, pp. 1–10, Feb. 2016, doi: 10.1038/srep20952.

[12] J. Lee, A. Seko, K. Shitara, K. Nakayama, and I. Tanaka, "Prediction model of band gap for inorganic compounds by combination of density functional theory calculations and machine learning techniques," *Phys. Rev. B*, vol. 93, no. 11, pp. 115104-115116, Mar. 2016, doi: 10.1103/PhysRevB.93.115104.

[13] J. Voss, "Machine learning for accuracy in density functional approximations," *J. Comput. Chem.*, vol. 45, no. 21, pp. 1829–1845, Oct. 2024, doi: 10.1002/jcc.27366.

[14] K. M. Tolle, D. S. W. Tansley, and A. J. G. Hey, "The fourth paradigm: Data-intensive scientific discovery," *Proc. IEEE*, vol. 99, no. 8, pp. 1334–1337, Aug. 2011, doi: 10.1109/JPROC.2011.2155130.

[15] M. Banerji, K. W. Willett, C. J. Lintott, S. P. Bamford, N. F. Cardamone, and A. S. Kaviraj, "Galaxy Zoo: Reproducing galaxy morphologies via machine learning," *Mon. Not. R. Astron. Soc.*, vol. 406, no. 1, pp. 342–353, Jun. 2010, doi: 10.1111/j.1365-2966.2010.16713.x.

[16] A. N. Jain and A. Nicholls, "Recommendations for evaluation of computational methods," *J. Comput. Aided Mol. Des.*, vol. 22, no. 3–4, pp. 133–139, Mar. 2008, doi: 10.1007/s10822-008-9196-5.

[17] T. CMS Collaboration, "Evidence for the direct decay of the 125 GeV Higgs boson to fermions," *Nat. Phys.*, vol. 10, no. 8, pp. 557–560, Jun. 2014, doi: 10.1038/nphys3005.

[18] A. White, "The materials genome initiative: One year on," *MRS Bull.*, vol. 37, no. 8, pp. 715–716, Aug. 2012, doi: 10.1557/mrs.2012.194.

[19] PyCaret, "Home - PyCaret," Accessed: Jan. 12, 2024. [Online]. Available: https://pycaret.org/

[20] E. T. Chenebuah, M. Nganbe, and A. B. Tchagang, "Comparative analysis of machine learning approaches on the prediction of the electronic properties of perovskites: A case study of ABX_3 and $A_2BB'X_6$," *Mater. Today Commun.*, vol. 27, no. 1, pp. 1–10, Jan. 2021, doi: 10.1016/j.mtcomm.2021.102462.

[21] F. Faber, A. Lindmaa, O. A. von Lilienfeld, and R. Armiento, "Machine learning energies of 2 million elpasolite crystals," *Phys. Rev. Lett.*, vol. 117, no. 13, pp. 135502-135514, Sep. 2016, doi: 10.1103/PhysRevLett.117.135502.

[22] G. Pilania and X. Y. Liu, "Machine learning properties of binary wurtzite superlattices," *J. Mater. Sci.*, vol. 53, no. 9, pp. 6652–6664, May 2018, doi: 10.1007/s10853-018-1987-z.

[23] Y. Mao, T. Ouyang, Y. Xie, M. Wu, and Y. Wang, "Prediction and classification of formation energies of binary compounds by machine learning: An approach without crystal structure information," *ACS Omega*, vol. 6, no. 22, pp. 14533–14541, Jun. 2021, doi: 10.1021/acsomega.1c01517.

[24] Nomad2018 Predicting Transparent Conductors | Kaggle, Accessed: Jan. 17, 2024. [Online]. Available: https://www.kaggle.com/c/nomad2018-predict-transparent-conductors

[25] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Comput. Sci.*, vol. 7, no. 1, pp. e623-e636, Mar. 2021, doi: 10.7717/peerj-cs.623.

[26] A. K. Gupta and M. Taufik, "Investigation of dimensional accuracy of material extrusion build parts using mathematical modelling and artificial neural network," *Int. J. Interact. Des. Manuf.*, vol. 17, no. 2, pp. 869–885, Apr. 2023, doi: 10.1007/s12008-022-01186-4.

[27] S. A. Shetty, T. Padmashree, B. M. Sagar, and N. K. Cauvery, "Performance analysis on machine learning algorithms with deep learning model for crop yield prediction," in *Proc. Int. Conf. Data Eng. Commun. Technol.*, vol. 2021, no. 1, pp. 739–750, Jan. 2021, doi: 10.1007/978-981-15-8530-2_58.

[28] M. A. Haque, R. Islam, M. N. Islam, F. Amin, and S. Adiba, "Machine learning-based technique for resonance and directivity prediction of UMTS LTE band quasi Yagi antenna," *Heliyon*, vol. 9, no. 9, pp. e19548-e19559, Sep. 2023, doi: 10.1016/j.heliyon.2023.e19548.

[29] Doreswamy, K. S. Harishkumar, Y. Km, and I. Gad, "Forecasting air pollution particulate matter (PM2.5) using machine learning regression models," *Procedia Comput. Sci.*, vol. 171, no. 5, pp. 2057–2066, May 2020, doi: 10.1016/j.procs.2020.04.221.

[30] C. Kittel, *Introduction to Solid State Physics*, 8th ed. New York, NY, USA: Wiley, 2005. [Online]. Available: http://103.62.146.201:8081/jspui/handle/1/9080

[31] D. L. Mohr, W. J. Wilson, and R. J. Freund, *Statistical Methods*, 4th ed. Cambridge, MA, USA: Elsevier, 2021, doi: 10.1016/B978-0-12-823043-5.00015-1.

[32] M. A. Uddin, M. Debnath, S. Roy, S. Adiba, and M. M. A. Talukder, "Identifying the smoking and smokeless tobacco-related predictors on frequencies of heavy vehicle traffic accidents in Bangladesh: Linear and binary logistic regression-based approach," *Adv. Civ. Eng.*, vol. 2023, no. 1, pp. 1-10, Jan. 2023, doi: 10.1155/2023/7116057.

[33] C. Nyasulu, A. Diattara, A. Traore, A. Deme, and C. Ba, "Exploring use of machine learning regressors for daily rainfall prediction in the Sahel region: A case study of Matam, Senegal," *Lecture Notes Inst. Comput. Sci., Soc.-Inform. Telecommun. Eng.*, vol. 459, no. may, pp. 78–92, May 2023, doi: 10.1007/978-3-031-25271-6_5.

[34] V. Rengaraj, R. Shankar, P. Jayaseelan, A. Parthasarathy, and N. Sakthipandi, "A two-step machine learning method for predicting the formation energy of ternary compounds," *Computation*, vol. 11, no. 5, p. 95, May 2023, doi: 10.3390/computation11050095.

[35] C. J. Bartel, A. Trewartha, Q. Wang, A. Dunn, A. Jain, and G. Ceder, "A critical examination of compound stability predictions from machine-learned formation energies," *npj Comput. Mater.*, vol. 6, no. 1, pp. 1–11, Jul. 2020, doi: 10.1038/s41524-020-00362-y.