BIBECHANA
Vol. 20, No. 3, December 2023, 267-274

ISSN 2091-0762 (Print), 2382-5340 (Online)
Journal homepage: http://nepjol.info/index.php/BIBECHANA
Publisher: Dept. of Phys., Mahendra Morang A. M. Campus (Tribhuvan University) Biratnagar

Machine learning driven prediction of lattice constants in transition metal dichalcogenides

Bhupendra Sharma[1], Laxman Chaudhary[1], Rajendra Adhikari[2] Madhav Prasad Ghimire[1,*]

[1]*Central Department of Physics, Tribhuvan University, Kirtipur, 44613, Kathmandu, Nepal.*
[2]*Department of Physics, Kathmandu University, Dhulikhel, Kavre, Nepal.*
[*]Corresponding author. Email: madhav.ghimire@cdp.tu.edu.np

Abstract

Machine learning represents an emerging branch of artificial intelligence, centering on the enhancement of algorithms in computer programs through the utilization of data and the accumulation of research-driven knowledge. The requirement for artificial intelligence in materials science is essential due to the significant need for innovative high-performance materials on a large scale. In this report, the gradient boosting regression tree model of machine learning was applied to predict the lattice constants of cubic and trigonal MX_2 systems (M=transition metal and X=chalcogen atoms). The theoretical/experimental values of the materials were compared to the predicted values to calculate the standard errors such as RMSE (root mean square error) and MAE (mean absolute error). The features used to predict lattice constants were ionic radius, lattice angles, bandgap, formation energy, total magnetic moment, density and oxidation states. The features versus contribution barplot has been drawn to reveal the contribution level of each parameter in the degree of [0,1] to obtain the predictions. This report provides a precise account of the prediction methodology for lattice parameters of the transition metal dichalcogenides family, a process that was previously not reported.

Keywords
Machine learning, Artificial Intelligence, Gradient Boosting Regression, Gradient Descent, RMSE, MAE.

Article information
Manuscript received: August 18, 2023; Accepted: September 23, 2023
DOI https://doi.org/10.3126/bibechana.v20i3.57732

1 Introduction

The crucial advancement of human history has been manifested by the development of new materials. So a vast amount of data has been collected throughout the centuries in the discipline of material science. The database of the previous experiments can be used to expedite innovations. The rise of artificial intelligence (AI) ushers a new genesis in the development of materials. AI works on the basis of complex multilayer neural networks with magnificent data mining ability. The fusion of material science and AI methods are used to find the complex relationship between different parameters,

predict the particular properties of materials and improve the material characterization techniques. The most favorable branch of AI in material science is Machine learning (ML) [1].

Recently, ML is turning out into a robust approach to study the materials with extremely fast and economical means. It works on the basis of neural networks (NN) which performs the prediction of required variables on the basis of training data. The first principle calculation of materials could now be enhanced by ML. We specifically design ML models to generate predictions based on the correlations of the database through statistical and probabilistic methods [2].

On the basis of nature of data labeling in material science, ML can be divided into two categories: Supervised learning and Unsupervised learning. Supervised learning is a ML approach which is used to evaluate an unknown mapping from known samples where the output is labeled (examples are classification and regression). This is like 'y' (dependent variable) is predicted with the knowledge of 'x' (independent variable) trained previously [3]. In classification, there are p predictor variables $x_1, x_2 ... x_p$ on which takes values 1, 2, ..., k are trained. Our motive is to find a model to predict the values of 'y' from new values of 'x'. The classification tree algorithm is used to identify the categorical target variable of the most probable "class" [4]. Also in regression models, the dependent variable is estimated due to the range of independent variable . The general linear regression model is:
$y = \beta_0 + \beta_1 x + \varepsilon$
where y is the dependent variable and x is independent variable. β_0 is the constant term which is the intercept of the regression line on the vertical axis. β_1 is regression coefficient which is actually the slope of the regression line. ε is the random error which can be used to express the effect of random factors on x [5].

Meanwhile, in unsupervised learning only input samples are provided to the learning system to make cluster and to estimate probability density function. Hence, the main goal of this learning system is to analyze the data, recognize a pattern and finally structure within the available set of data.

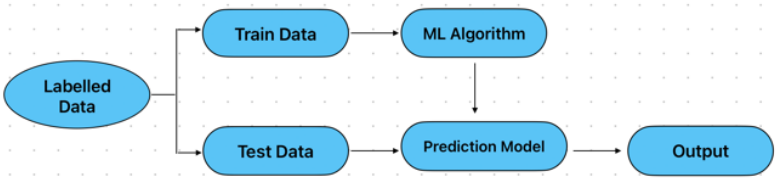

Figure 1: Basic architecture of supervised learning.

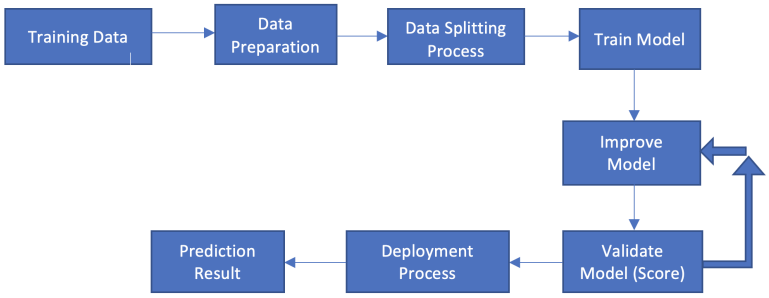

Figure 2: Schematic picture of stepwise machine learning.

1.1 Process of Machine Learning

The major steps involved in the process of ML are: [6]

- Identify the problem you want to solve using ML
- Collection of training data for the process
- Selection of ML model
- Preparation of accumulated data to train in ML model
- Test your ML system with test data

- Validation and improvement of the ML model. Usually, we need to search more training data during this iterative loop.

1.2 Density Functional Theory to Machine Learning

Recently, we have entered the fourth paradigm of science called data-driven science led by the experiments and simulations. The influence of this paradigm has led to the rise of new field called material informatics. The objective of material informatics is to discover the relationship between known standard features and materials properties such as structure, symmetry, composition, and properties of the constituent elements. The imperceptible feature property relationships beyond human capacities are recognized through ML process [7].

A combination of density functional theory and machine learning techniques provide a practical method to excavate feature-property relationships much more efficiently than by DFT or experiments [8]. As the fundamental mathematical model of DFT works only for ground state density so the study of excited states is limited within this method. The strongly correlated systems, such as d-electron in transition metal oxides (TMO) are solved with supplemental theories like Hubbard parameter (U). To extend the proficiency of DFT, various auxiliary code based models are integrated with DFT codes such as VASP, Quantum Espresso, and WIEN2k. These methods are vastly computationally demanding but consistent and reliable. Hence, the abundant database obtained from successful DFT platforms can be fitted into appropriate ML models to steer the discovery of complex properties of materials [7].

1.3 Transition Metal Dichalcogenides

Two dimensional transition metal dichalcogenides (2D-TMDs) are layered materials with robust in-plane bonding and weak van der Waals interactions between the planes which enables the exfoliation into thickness of two dimensional layers with single unit cell [9]. TMDs are an emerging class of materials with highly attractive properties such as atomic-scale of thickness, direct bandgap and strong spin orbit coupling. Hence these materials accelerate the fundamental studies of novel physical phenomena with applications ranging from nano-electronics and nanophotonics to sensing and stimulation at the nanoscale [10]. The different properties of bulk TMDs ranges from insulators such as HfS_2, semiconductors such as MoS_2 and Ws_2, semimetals such as WTe_2 and $TiSe_2$, to true metals such as NbS_2 and VSe_2. The properties are mostly preserved with the exfoliation of such materials into mono or few layers with supplemental characteristics that emerges from confinement [11].

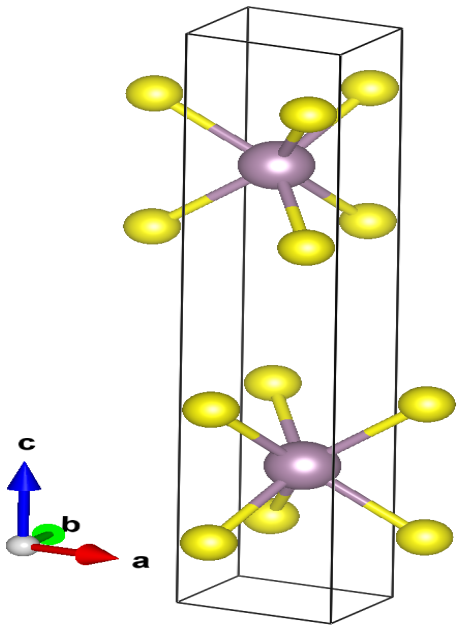

Figure 3: Bulk crystal structure of MoS_2.

2 Materials and Methodology

In this section, the methodology and the required programming model used in our work has been described.

2.1 Data Mining

It is necessary to collect qualitative data for researchers to create the model for specific problems. Open source databases such as Materials Project [12] and Open Quantum Materials Database [13] are one of reliable and representative databases. These databases could be used to extract the data set of the materials of our concern.

Once the collection of data is done, we need to clean the problems in data such as the data containing abnormal or redundant values. This is called data cleaning. The data cleaning steps are:

- Data Sampling
- Processing of abnormal value
- Discretization of data
- Data Normalization

Data sampling allows us to create prediction model of high performance with less data. Thereafter, the removal of abnormal values maintains the accuracy of the model. The continuous features are reduced by data discretization. Finally, data normalization is used to organize the columns and tables on the basis of dependencies seen in data. After these steps, the data set is split into training and testing sets to subject it into the ML model [14].

2.2 Gradient Descent

The objective of Gradient Descent Algorithm is to minimize a differentiable cost function in certain number of iterations [15]. If $y = mx + c$ is regression model then m and c are parameters. So, to find the optimum value of y, m (slope) and c (y-intercept) must be minimized. Then the cost function becomes:

$$J(m,c) = \sum_{i-1}^{n} (y_i - (mx + c))^2 \tag{1}$$

This cost function is used to optimize the parameters m and c with following relation.

$$m_{\text{new}} = m_{\text{old}} - \rho \frac{\partial J(m,c)}{\partial m}\bigg|_{(m',c')} \tag{2}$$

where, ρ is the learning rate and (m', c') is initial guess. Similarly,

$$c_{new} = c_{cold} - \rho \frac{\partial J(m,c)}{\partial m}\bigg|_{(m',c')} \tag{3}$$

So the cost function has to be minimized by moving in opposite direction of the gradient. The iteration process will be repeated until we reach the minimum cost function. Finally, we obtain the optimized parameters for best fit in our regression model [16].

2.3 Cross-Validation

Cross-validation technique is a resampling method in ML where the observation set is split into two subsets. The first subset is called training set and the second one is called test set. The training set is exercised to find a proper function whose predictive capacity is based on the prediction error. Prediction error is calculated by applying the result of training set into the test set.

In this report, we have used k-fold cross validation method because it estimates the model in finer way by training and evaluating it k-times enhancing the model performance. In this method, the dataset is partitioned into k subsets (folds), then the function is trained on the k-1 subsets. Finally, the prediction error is estimated on each k-1 subsets with the help of the remaining set [17]. The other validation methods such as Holdout validation that splits whole data into training and validation set could be sensitive to the initial random split and the evaluation can biased depending on the split. Also, we opt k-fold cross-validation as our preferred method for model evaluation as it is most suitable for smaller datasets. Ultimately, the choice of validation technique is chosen on the basis of the characteristics of our data and the research objectives.

2.4 Gradient Boosting Regression Model

We have used Gradient Boosting regression tree (GBRT) to predict the lattice parameters of cubic and trigonal TMDs. A GBRT model minimizes the prediction errors through boosting technique which compiles the set of weak models to construct a single strong model. In this technique, the prediction error is minimized sequentially by generating new decision trees. Decision tree compares the values of root attribute with the real data-set attribute and on the basis of the comparison, it follows the branch and moves to the next node. The sequential process is basically a functional gradient descent where the prediction is optimized with the addition of new tree at every step, in order to minimize the cost function [18].

In this work, we have predicted the lattice constants (*a, b* and c) using other properties of MX_2 system such as ionic radius, lattice angles, band gap, formation energy, total magnetic moment, density and oxidation number using GBRT model. The hyperparameters of GBRT model used here are

max-depth and n-estimators. The max-depth sets the maximum depth of nodes in a tree which is set to 4. The n-estimator chooses the number of trees to be considered in our model which is set to be 500. The 3-fold cross validation is applied to the model with shuffle on mode. This means that the dataset has been split into three subsets. Two of them were used for training where remaining one was used for testing the model and the data are trained and tested in shuffle mode.

3 Results and Discussion

The interpretation of the accuracy of our model is measured on the basis of Root mean square error (RMSE) and mean absolute error (MAE). For n observations y with n corresponding model predictions $\hat{y}$ (y_i, i=1,2...n), RMSE and MAE modeled by Hodson [19] are as below:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{5}$$

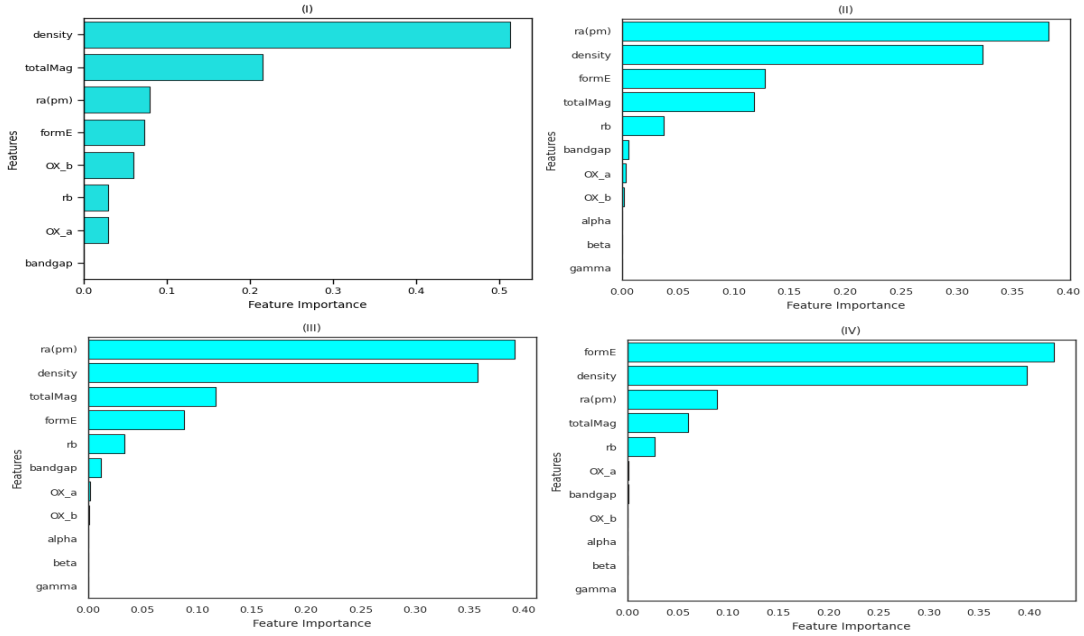

Figure 4: Barplot (i) shows the features used and their roles in the model to predict a of the cubic MX_2 system (ii) shows the features used and their roles in the model to predict a of trigonal MX_2 system (iii) shows the features used and their roles in the model to predict b of trigonal MX_2 system (iv) shows the features used and their roles in the model to predict c of trigonal MX_2 systems. The figures were generated in Jupyter Notebook using Python codes.

3.1 Lattice Constants Prediction for Cubic Systems

Twenty one cubic TMDs that were available in materials project website has been used. We used GBRT algorithm to predict the lattice constant of such system. Since in cubic system $a = b = c$ and $\alpha = \beta = \gamma$, so it is sufficient to predict one lattice constant for this structure.

Average RMSE and MAE from three validations are 2.62 Å and 0.824 Å, respectively. Figure 4 (i) shows that the order of importance of the parameters is density, total magnetic moment, ionic radius of first element, formation energy, oxidation state of second element, ionic radius of second element, oxidation state of first element, and bandgap (denoted as density, totalMag, ra(pm), formE, OX_b, r_b, OX_a, and band gap, respectively, in Figure 4. So it is seen that the density has the most important role in pattern recognition by the model, and band gap has the least. This is because almost all data of cubic systems possess zero band gap.

Table 1: Comparison table of predicted lattice constants versus their experimental/theoretical values in cubic MX_2 system.

Material	Experimental/Theoretical (a)	Predicted (a')
$MnTe_2$	6.69	6.17
CoS_2	5.51	5.58
$CoSe_2$	5.84	5.89
NiS_2	5.61	5.71
$NiSe_2$	5.94	6.13
YS_2	7.83	5.87
$CuSe_2$	6.16	6.31

3.2 Lattice Constants Prediction for Trigonal Systems

Thirty-one trigonal TMDCs have been taken from project materials to predict the lattice constants of such systems using a GBRT model. In the trigonal crystal system, the relation between major lattice parameters (lattice constants and lattice angles) is given by $a = b \neq c$, $\alpha = \beta = 90^0$ and $\gamma = 120^0$. The calculated predictions of lattice constants are shown in Table 2.

Figure 4(ii) shows that the prediction of a is mostly influenced by the ionic radius to draw the pattern relation, whereas the lattice angles have the least important role, primarily because the angles of all the systems are identical. The average standard errors after three validations are: RMSE = 0.0542 Å and MAE = 0.205 Å.

In Figure 4(ii) and (iii), a and b are equal for the trigonal system, so almost identical errors are seen in predicting b with the same feature roles.

In the prediction of c, formation energy has shown the most important role in pattern recognition, while density contribution is in second place, as shown in Figure 4(iv). The average standard errors after three validations are found to be RMSE = 0.547 Å and MAE = 0.462 Å.

Table 2: Comparison table of predicted lattice constants versus their experimental/theoretical values in trigonal MX_2 systems.

Material	a	a'	b	b'	c	c'
$FeTe_2$	3.75	3.75	3.75	3.66	5.84	5.23
VTe_2	3.66	3.64	3.66	3.55	6.95	6.13
$CoTe_2$	3.79	4.02	3.79	4.00	5.56	5.19
$CuTe_2$	4.02	3.90	4.02	3.90	5.10	5.24
$ZnTe_2$	4.19	3.83	4.19	3.83	5.20	6.12
$ZrTe_2$	3.98	3.79	3.98	3.79	7.00	6.73

4 Conclusions

In this work, the prediction of lattice constants of cubic and trigonal MX_2 systems has been done through a Gradient Boosting Regression model of machine learning. We have predicted a, b, and c on the basis of patterns or relationships drawn through training the algorithm with features such as ionic radius of atoms, lattice angles, bandgap, formation energy, total magnetic moment, density, and oxidation states. In the cubic system, the RMSE and MAE in the prediction of lattice constants were found to be 2.62 Å and 0.824 Å, respectively. In the trigonal system, the RMSE of $a = b$ and c were 0.0542 Å and 0.547 Å, respectively. Whereas, the MAE of $a = b$ and c were 0.205 Å and 0.462 Å, respectively.

The barplot of cubical systems shows that the contribution of density is the most significant in prediction, followed by total magnetic moment and ionic radius. Whereas, the barplot of the trigonal system indicates ionic radius as the major contributor in the prediction of $a = b$, and formation energy as the major contributing parameter to predict c.

DFT calculations and experiments are discovering a plethora of materials day by day. Hence, a large collection of data can be acquired in the field of 2D-TMDs in the near future. The consequence is that the huge database could be used in machine learning to accelerate materials discovery with higher precision in the least amount of time. Thereafter, not only lattice constants but every lat-

tice parameter we desire can be predicted with extreme accuracy.

Applications

The cubic and trigonal TMDs are one of most stable, extensively researched and then advantageous form of 2D materials. Predicting their lattice parameters using machine learning has numerous practical implications across materials science and related fields. These predictions can expedite materials discovery by guiding researchers toward promising TMD combinations, impacting areas like catalysis and energy storage. Moreover, the precise control offered by these predictions aids in designing TMDs for electronic and optical applications, influencing the development of semiconductors, photodetectors, and optoelectronic devices. This can lead to advancements in electronics, photonics, and telecommunications. They are also invaluable for materials characterization, aiding experimentalists in structural analysis techniques. Experimentalists can use these predictions as reference values during structural analysis techniques such as X-ray diffraction and electron microscopy to validate the quality of synthesized samples. Overall, machine learning's role in TMD lattice parameter prediction accelerates innovation, improves materials properties, and reduces costs across a broad spectrum of applications.

Ethical considerations

In our study, we adhere to ethical guidelines, even when utilizing data sourced from the Materials Project. Although the data is not confidential, we uphold ethical standards because our research can have societal implications. We adhere to the Materials Project's data privacy policies, and we take deliberate steps to prevent bias through the careful selection, analysis, and transparency of data. For additional information about the Materials Project, please visit the website [12].

Acknowledgments

This work was supported by a grant from UNESCO-TWAS and the Swedish International Development Cooperation Agency (SIDA) with TWAS Research Grant Award No. 21-377 RG/PHYS/AS-G. The views expressed herein do not necessarily represent those of UNESCO-TWAS, SIDA or its Board of Governors. Part of this work was also supported by the University Grants Commission, Nepal under UGC Grant No. CRG-78/79 S&T-03. We acknowledge NVIDIA for providing the license package of Deep learning, a part of machine learning.

References

[1] Wenqing Sha, Yike Guo, Qing Yuan, Shuai Tang, Xiaoyan Zhang, Shan Lu, Xiaojun Guo, Yang-Chun Cao, and Shijin Cheng. Artificial intelligence to power the future of materials science and engineering. *Advanced Intelligent Systems*, 2(4):1900143, 2020.

[2] Haoyuan Liang, Valentin Stanev, Aaron G Kusne, and Ichiro Takeuchi. Cryspnet: Crystal structure predictions via neural networks. *Physical Review Materials*, 4(12):123802, 2020.

[3] Issam El Naqa and Martin J Murphy. *What Is Machine Learning?* Springer, 2015.

[4] Wei-Yin Loh. Classification and regression trees. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, 1(1):14–23, 2011.

[5] Dauda Maulud and Abdulmalik Abdulazeez. A review on linear regression comprehensive in machine learning. *Journal of Applied Science and Technology Trends*, 1(4):140–147, 2020.

[6] J Ross Quinlan. *C4. 5: Programs for Machine Learning*. Elsevier, 2014.

[7] Gabriel Rodrigues Schleder, Antonio Carlos Padilha, Camila Maciel Acosta, Maurício Costa, and Adalberto Fazzio. From dft to machine learning: recent approaches to materials science–a review. *Journal of Physics: Materials*, 2(3):032001, 2019.

[8] Sherif A Tawfik, Olexandr Isayev, Catherine Stampfl, Joe Shapter, David A Winkler, and Michael J Ford. Efficient prediction of structural and electronic properties of hybrid 2d materials using complementary dft and machine learning approaches. *Advanced Theory and Simulations*, 2(1):1800128, 2019.

[9] Q H Wang, K Kalantar-Zadeh, A Kis, J N Coleman, and M S Strano. Electronics and optoelectronics of two-dimensional transition metal dichalcogenides. *Nature nanotechnology*, 7(11):699–712, 2012.

[10] S Manzeli, D Ovchinnikov, D Pasquier, O V Yazyev, and A Kis. 2d transition metal dichalcogenides. *Nature Reviews Materials*, 2(8):1–15, 2017.

[11] Manish Chhowalla, Hyeon Suk Shin, Goki Eda, Li-Jie Li, Kian Ping Loh, and Hua Zhang. The chemistry of two-dimensional layered transition metal dichalcogenide nanosheets. *Nature chemistry*, 5(4):263–275, 2013.

[12] https://next-gen.materialsproject.org/

[13] https://oqmd.org/:The open quantum materials database.

[14] J Cai, X Chu, K Xu, H Li, and J Wei. Machine learning-driven new material discovery. *Nanoscale Advances*, 2(8):3115–3130, 2020.

[15] Souvik Ray. A quick review of machine learning algorithms. In *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, pages 35–39. IEEE, 2019.

[16] A gentle introduction to gradient boosting, 2016.

[17] Tim Mueller, Aaron G Kusne, and Rampi Ramprasad. Machine learning in materials science: Recent progress and emerging applications. *Reviews in computational chemistry*, 29:186–273, 2016.

[18] Fan Yang, Dan Wang, Fan Xu, Zhitao Huang, and Kwok L Tsui. Lifespan prediction of lithium-ion batteries based on various extracted features and gradient boosting regression tree model. *Journal of Power Sources*, 476:228654, 2020.

[19] Timothy O Hodson. Root mean square error (rmse) or mean absolute error (mae): when to use them or not. *Geoscientific Model Development Discussions*, pages 1–10, 2022.