JNPS 10 (2): 79-88 (2024)
ISSN: 2392-473X(Print), 2738-9537 (Online)

Research Article *©Nepal Physical Society*
doi.org/10.3126/jnphyssoc.v10i2.79494

Implementation of Tree Based Machine Learning Model to Predict Stability of Multicomponent Materials of B and Si Using Compositional-based Features

Subash Dahal, Devendra Adhikari, Shashit Kumar Yadav*
Department of Physics, Mahendra Morang Adarsh Multiple Campus, Tribhuvan University, Biratnagar.
**Corresponding Author: yadavshashit@yahoo.com*

Abstract
The stability of materials plays a crucial role in expediting the development in material science and engineering. The stability factor determines the synthesizability of the material. The utilization of Density Functional Theory (DFT) enables the examination of material stability; however, the process is hindered by the high computational costs and time-consuming calculations, making it challenging to forecast the stability of numerous potential materials. In this work, a machine learning model (MLM) was developed to anticipate material stability by taking into account compositional-based features. A total of 8763 multicomponent materials of B and Si from the material project database were used in this study. The Elemental Property compositional-based featurizer from the MATMINER packages was employed to create 133 new features, out of which only 25 were chosen using the forward selection technique. The dataset was divided into training and testing sets with a test size of 0.2 for model training. The model was trained, and hyperparameters were fine-tuned through 10-fold cross-validation using the Scikit library in the Anaconda distribution. The Random Forest Classifier and Gradient Boosting Classifier exhibit noteworthy accuracies of 0.873 with F1 scores of 0.851 and 0.867, respectively, on the test data. Conversely, the Extra Trees Classifier demonstrates slightly inferior performance compared to the aforementioned models, yet it achieves a satisfactory F1 score of 0.80 and an accuracy of 0.849. Our investigation demonstrates the potential of machine learning models in predicting material stability thus aiding researchers in expediting material discoveries.

Keywords: Material Stability, ML, MATMINER, Density Functional Theory, Multicomponent Materials, Performance Metrics.

Submitted: November 12, 2024; **Revised:** May 30, 2025; Acceped: June 01, 2025

1. Introduction
Stability prediction of materials in materials science and engineering plays a pivotal role to accelerate the material discovery. It is important to enhance the robustness, mechanical properties, and longevity which contribute in reducing cost for maintenance, improve quality, and corrosion resistance of materials.

The diverse application requirements of material are also the outcomes of this material stability. The prediction of stability poses a challenge due to complex nature of material, thermodynamic factors, and kinetic factors. Moreover, stability prediction is complicated by mathematical principles and organizational complexities. Therefore, understanding of materials and mathematical principles is necessary for accurate stability prediction. To overcome the challenges associated, many researchers have utilized atomistic theories which provide a holistic approach to deal with the complexities inherent in predicting stability. Furthermore, the development of advanced computational methods, such as Density Functional Theory (DFT) with appropriate functionals can be used to predict the stability. Nevertheless, challenges persist on DFT due to demanding computational expenses [1]–[3].

Recently, machine learning models (MLM) have been used by different researcher in the field of material science. Gajera et al. in 2022 utilized a machine-learning (ML) technique to forecast the energetic stability of semiconducting binary compounds, specifically focusing on zinc blende and rocksalt crystal structures [4]. Their research indicates that a straight-forward 1D formula based on atomic attributes effectively captures the energetics, with spatial atomic characteristics playing a significant role. Furthermore, the study underscores the relevance of atomic size in determining compound energetics, illustrating the effectiveness of ML in anticipating compound stability [4]. Tawfik et al. in 2023 implemented a ML classifier on a dataset containing about 3100 materials to evaluate vibrational stability [5]. Their investigation demonstrates that ML classifier exhibits the ability to accurately distinguish between stable and unstable materials. Rengaraj et al. in the same year proposed a two-step machine learning methodology for forecasting the formation energy of ternary compounds [6]. Their outcomes revealed that the machine learning model adeptly predicts the formation energy of ternary compounds. Additionally, they predicted that the incorporation of a centralized Adam optimizer enhanced prediction accuracy, thereby bolstering the model's performance. Tanabe employed a ML-based strategy to predict the stability of V-Cr-Ti alloys, which aligned well with experimental findings on the ductile-brittle transition temperature and swelling behavior [7]. Their findings underscored the usefulness of machine learning in crafting alloys, particularly for applications in nuclear reactors. In general, the predictive capabilities of ML models highlighted its potential in designing multicomponent alloys for nuclear fusion reactors [7]. Datta et al. utilized a Support Vector Regression (SVR) model to estimate the stability of silicon (Si)-alkaline metal alloys, emphasizing its adaptability to new Si alloys with unique electronic configurations and structures [8]. Their work showcased the model's utility in novel Si alloy compositions and highlighted the importance of hyperparameter tuning and training data selection in ML. Hong et al. employed ML models for predicting the crystal structure using dynamical trajectories based on density functional theory [9]. Ihalage et al. used 55, 11, 10 and 5 elemental, compositional, SISSO, and

combination of t and μ features respectively to find the fingerprints of disordered Perovskites [10]. In our previous work, we implemented the compositional based features only in dataset containing 973 rows to predict the formation energy of copper based ternary alloys where Gradient Boosting Regressor explain the 94% of the variance of the dataset [11].

Silicon-based alloys offer significant advantages due to their cost-effectiveness, durability, and suitability for use in a variety of environmental conditions. The inclusion of aluminum in silicon alloys enhances their strength, making them ideal for applications requiring exceptional toughness. Moreover, the incorporation of silicon-based alloys in batteries enhances their performance in electronic devices, increases energy storage capacity, and extends their lifespan [12]–[16]. Similarly, the use of boron alloy provides improved corrosion resistance, offers a cost-effective alternative, enhances mechanical properties, and enables efficient mass production. The straightforward preparation process and versatile characteristics of boron alloy make it well-suited for a wide range of applications across various industries [17]–[20].

The main goal of this research is to utilize a combination of machine learning algorithms such as RandomForestClassifier, GradientBoostingClassifier, and ExtraTreesClassifier on a dataset obtained from a material project database. While descriptors such as XRD, SEM, phonon band gap, formation energy, and structural information can be utilized for stability classification, this study focuses on classifying the stability of multicomponent alloys composed of boron and silicon using only compositional features extracted from MATMINER [21]. The choice to rely exclusively on compositional descriptors is motivated by the challenges associated with obtaining the aforementioned descriptor data for all materials.

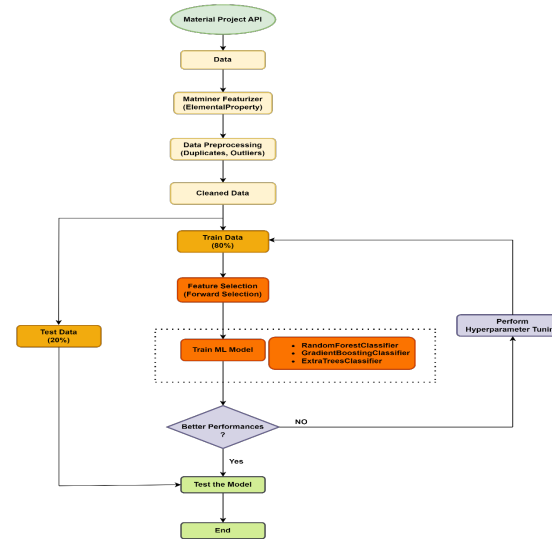


Figure 1: Different steps involved in developing Machine Learning algorithm to predict material stability.

2. Methodology

The construction and validation of machine learning model requires several steps: (i) Data collection (ii) Data cleaning and Feature generation (iii) Feature selection (iv) Hyperparameter tuning and model selection (v) Model validation. Figure 1 shows the different steps involve while preparing the machine learning model to predict the stability of materials. All these steps are performed using scikit learn [5], [22].

2.1 Data collection

The data utilized in our study was obtained from the material project database, which was accessed by using an API key provided by the database [23]. This dataset comprises 8763 chemical systems that provide information on material Id, chemical formula, composition, and stability of chemical systems. These systems involve 4 and 5 components chemical system of B, as well as 4 and 6 components chemical system of Si. The chemical systems are denoted as [B-*-*-*, B-*-*-*-*-*, Si-*-*-*, Si-*-*-*-*-*], with each * representing a distinct chemical species.

2.2 Data cleaning and feature generation

The collected data undergoes a verification process for duplicates, which are subsequently eliminated before proceeding to the creation of features. Utilizing the Elemental Property Magpie featurizer from MATMINER, we were able to generate a total of 133 novel features. These newly created features encompass a variety of information related to distinct attributes, including atomic number, atomic mass, space group, etc. Subsequently, we employed the Isolation Forest model with a contamination rate of 0.1 and a random state of 42 to identify any outliers present in the data. A total of 707 outliers were identified and subsequently excluded from the dataset. Figure 2 shows the total number of stable and unstable chemical system present in cleaned dataset. The preprocessed dataset was then partitioned into training and testing sets, with a test size of 0.2.

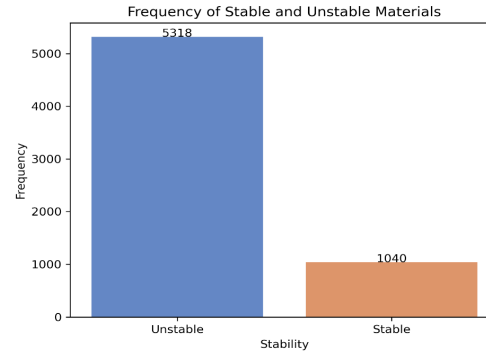


Figure 2: Frequency of stable and unstable chemical systems.

2.3 Feature selection

The training data and testing data were standardized using the StandarScaler class with its default parameters from the scikit learn library. The standardized training data were subsequently employed for model development. To conduct feature selection, we applied the sequential feature selection technique utilizing the SequentialFeatureSelector class in the mlxtend package [24]. It is a collection of Python tools and extensions designed for Machine Learning tasks. Within the SequentialFeatureSelector, we utilized the RandomForestClassifier model and executed forward selection to determine the optimal number of features. Subsequently, it is found that 25 features represent the most suitable choice for the machine learning model to operate at its peak performance level.

2.4 Hyperparameter tuning and model selection

Hyperparameters are the parameters in ML that control the learning process of the machine learning model. We used

GridSearchCV from the Scikit Learn library to identify the best hyperparameter for our model. We used 10-fold cross-validation in GridSearchCV to identify the best hyperparameter for our models. Table 1 shows the different hyperparameters used and the best hyperparameters obtained with the help of GridSearchCV for our machine learning models. On the other hand, to identify the best model for our problem, we used accuracy, precision, recall, and the F1 score as our performance metrics. The proportion of accurate predictions made by a machine learning model is known as accuracy. Precision is the proportion of predicted positives that are actually positive. Recall, on the other hand, is the proportion of actual positives that are correctly classified. The F1 score is the harmonic mean of precision and recall. The F_1 score is used here to overcome the problem of selecting the Type I or Type II error as it balances the precision and recall. The mathematical equation for these metrics is as follows

$$Accuracy = \frac{\text{True Positive} + \text{True Negative}}{\text{Total Classification}} \quad (1)$$

$$Precision = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \quad (2)$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \quad (3)$$

$$F_1 \text{ score} = \frac{2 \times \text{precision} \times \text{Recall}}{\text{precision} + \text{Recall}} \quad (4)$$

2.5 Model validation

Using the optimum parameters found through the hyperparameter tuning method, we first trained the three models, Random Forest Classifier, Gradient Boosting

Table 1: Hyperparameter used and obtained best parameter in different model

Model	Parameter used	Best parameters
RandomForestClassifier	bootstrap: [True, False], max_depth: [1,2,3,4,5,6,7], min_samples_leaf: [1,2,3,4,5,6,7], max_leaf_nodes: [None, 5, 10, 20, 50], max_samples: [0.5, 0.75, 0.85], n_estimators: [10,20,30,40,50,100,150,200]	bootstrap: True, max_depth: 7, max_leaf_nodes: None, max_samples: 0.85, min_samples_leaf: 1, n_estimators: 30
GradientBoostingClassifier	n_estimators: [10,30,50,70, 100, 150], max_depth: [1,3, 5, 7], min_samples_split: [2, 5, 7,10], min_samples_leaf: [1, 2, 3,4]	max_depth: 7, min_samples_leaf: 1, min_samples_split: 2, n_estimators: 150
ExtraTreeClassifier	n_estimators: [10,30,50,70, 100, 150], max_depth: [1,3, 5, 7], min_samples_split: [2, 5, 7,10], min_samples_leaf: [1, 2, 3,4]	max_depth: 7, min_samples_leaf: 1, min_samples_split: 7, n_estimators: 70

Classifier, and ExtraTreesClassifier, in order to verify the models. Next, we validated the model using our test data, and based on the performance metrics mentioned in the Section 2.4, then after the best model was chosen.

3 Results and discussion

3.1 Feature selection

The impact of the number of features on the performance of a machine learning model was examined through the methodology outlined in Section 2.3. The graphical representation in Figure 3 illustrates the influence of the number of features on model performance, demonstrating an initial sharp increase in performance with the inclusion of a few key features. It was observed that the model's performance reaches its peak at 25 features, after which a decline in performance ensues, leading to a plateau effect. Consequently, the number of features was reduced from 133, generated using Matminer to 25. This reduction emphasizes the balance achieved between model complexity and performance, as an excessive number of features could potentially lead to model overfitting. Moreover, Figure 4 illustrates the importance of features attributed to our machine learning model in forecasting the stability of materials within our dataset. Our investigation revealed that this set of 25 features plays a crucial role in stability prediction.

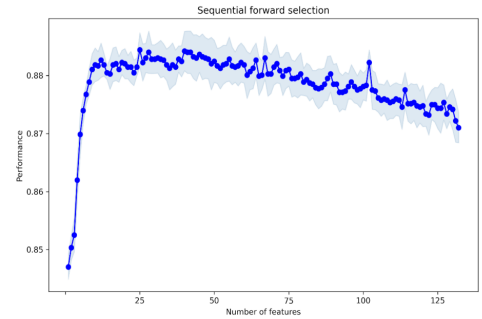


Figure 3: Plot of number of features versus performance.

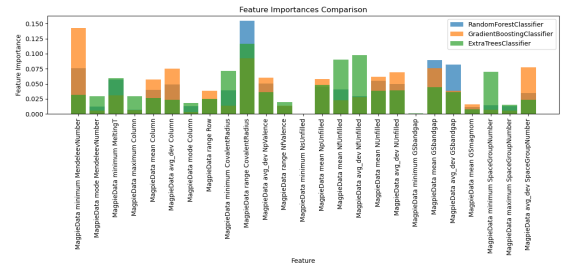
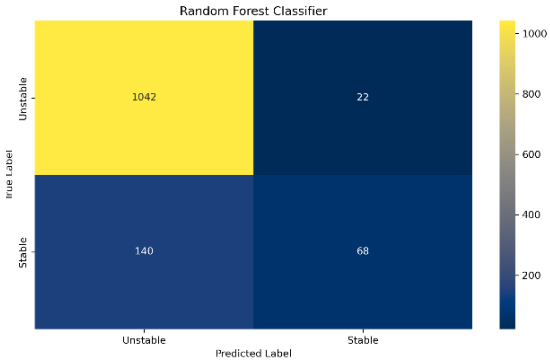


Figure 4: Feature importance of different features in Machine learning models.

3.2 Classification of stable/unstable materials

Three distinct machine learning models, namely Random Forest Classifier, Gradient Boosting Classifier, and Extra Trees Classifier, were applied to categorize materials as stable or unstable after training the model with hyperparameter presented in Table 1. The confusion matrix illustrating the classification performance of these models is presented in Figure 5. Test datasets consisting of 1272 (1064 unstable and 208 stable) multicomponent materials of B and Si materials were utilized to construct these confusion matrices. The RandomForestClassifier identified 1042 unstable materials correctly and 68 stable materials accurately. Nevertheless, 140

materials were classified as unstable by the model despite being stable, while 22 materials were classified as stable despite being unstable. Consequently, out of 1272 predictions, 1110 were deemed accurate by the RandomForestClassifier model. Similarly, the GradientBoostingClassifier achieved a total of 1111 accurate predictions, with 105 being stable materials and 1006 being unstable materials. However, 161 predictions made by the GradientBoostingClassifier were inaccurate, including 103 materials predicted as unstable that were actually stable and 58 materials predicted as stable that were actually unstable. Moreover, the ExtraTreesClassifier yielded 1080 correct predictions, with 1055 pertaining to unstable materials and 25 to stable materials. This model also produced 112 incorrect predictions, with 103 stable materials inaccurately classified as unstable and 9 unstable materials inaccurately classified as stable. From this information it is clear that in terms of correct prediction, GradientBoostingClassifier performed well while ExtraTreesClassifier performed the least out of these three models.



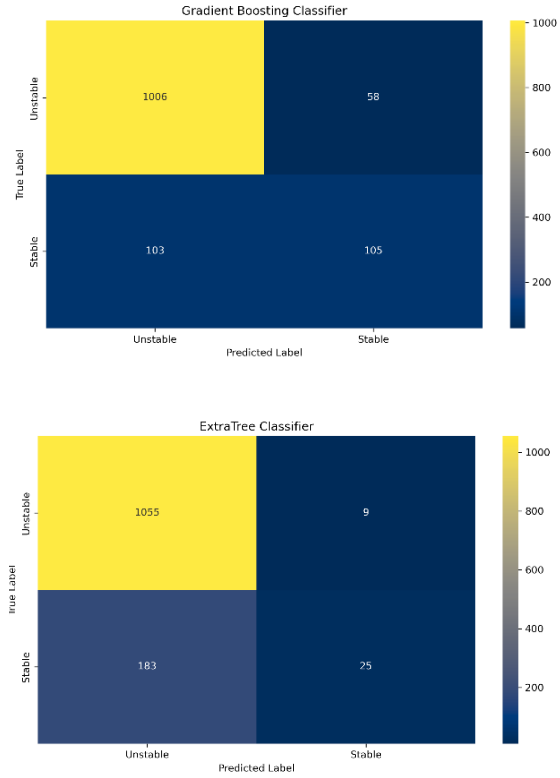


Figure 5: Confusion matrix for different models.

3.3 Performances of different models

The accuracy, precision, recall, and F1 score of all the three models are directly obtained from the Scikit-Learn library utilizing the data provided in section 3.2 and the mathematical equations outlined in section 2.4. These metrics for each model are illustrated in Table 2. Random Forest Classifier and Gradient Boosting Classifier demonstrate robust performance, boasting precision scores of 0.861 and 0.864 respectively, each with identical recall scores of 0.873, and F1 scores of 0.851 and 0.867 respectively. These classifiers also exhibit an accuracy of 0.873, indicating their capability in predicting material stability.

Conversely, the ExtraTreesClassifier displays slightly inferior results compared to RandomForestClassifier and GradientBoostingClassifier, with precision of 0.735, recall of 0.849, F_1 score of 0.80, and accuracy of 0.849. Despite its lower performance relative to the other two models, it showcases reasonably good results, especially in terms of recall. All these details are depicted in Figure 6. In conclusion, our investigation highlights the suitability of RandomForestClassifier and GradientBoostingClassifier for predicting material stability, while recognizing the satisfactory performance of ExtraTreesClassifier for this predictive task. Moreover, our research emphasizes the significance of machine learning models in forecasting material stability exclusively based on compositional features.

Table 2: Performance metrices for all three models

	Model		
	RandomForestClassifier	GradientBoostingClassifier	Extra TreesClassifier
Precision	0.861	0.864	0.735
Recall	0.873	0.873	0.849
F_1 Score	0.851	0.867	0.800
Accuracy	0.873	0.873	0.849

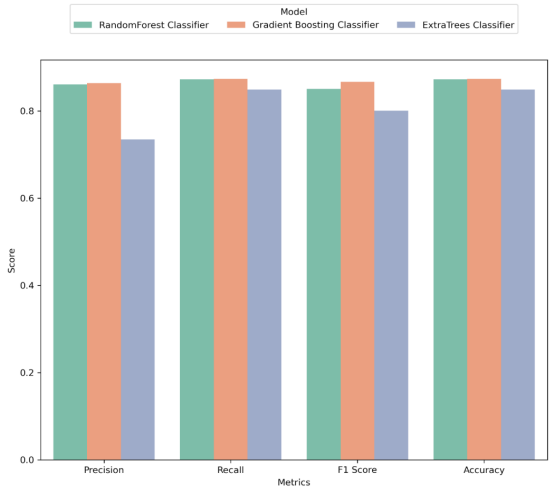


Figure 6: Performance metrics comparison of all three models.

4. Conclusion

In this work, three machine learning models (Random Forest Classifier, Gradient Boosting Classifier, and Extra Trees Classifier) were employed to forecast the stability of multicomponent materials of B and Si. The dataset utilized initially encompassed 8763 materials obtained from the material project database. Subsequently, a total of 133 novel compositional based features were extracted using MATMINER, from which a subset of 25 optimal features was identified through the forward selection technique. Following this feature selection process, the dataset was partitioned into training and testing sets, with a test size ratio of 0.2. The machine learning model was then constructed using the training data, involving hyperparameter optimization through 10-fold cross-validation. Evaluation of the models were conducted using the test dataset, demonstrating notable performance metrics including precision, recall, F_1

score, and accuracy. The results indicate that compositional features alone provide meaningful insights into material stability, with the GradientBoostingClassifier achieving the highest overall performance in terms of precision (0.864), recall (0.873), F1-score (0.867), and accuracy (0.873). While additional descriptors such as XRD, SEM, phonon band gap, and formation energy could enhance predictive accuracy. But this study demonstrates that compositional features offer a viable alternative for stability classification, particularly when experimental or computational data is unavailable. These findings highlight the potential of machine learning in accelerating the discovery of stable multicomponent materials based solely on composition. Furthermore, the findings of this investigation suggest that tree-based machine learning models exhibit potential in elucidating material stability based on compositional based features.

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work presented in this study.

References

[1] R. Pascal and A. Pross, "The nature and mathematical basis for material stability in the chemical and biological worlds," 2014.

[2] L. Zhang, "Stability analysis of atomic structures," University of Iowa, 2006.

[3] Y. Zhang *et al.*, "Efficient first-principles prediction of solid stability: Towards chemical accuracy," *NPJ Comput Mater*, vol. 4, no. 1, p. 9, Dec. 2018.

[4] U. K. Gajera, L. Storchi, D. Amoroso, F. Delodovici, and S. Picozzi, "Towards machine learning for microscopic mechanisms: a formula search for crystal structure stability based on atomic properties," Feb. 2022.

[5] S. A. Tawfik, M. Rashid, S. Gupta, S. P. Russo, T. R. Walsh, and S. Venkatesh, "Machine learning-based discovery of vibrationally stable materials," *NPJ Comput Mater*, vol. 9, no. 1, Dec. 2023.

[6] V. Rengaraj *et al.*, "A Two-Step Machine Learning Method for Predicting the Formation Energy of Ternary Compounds," *Computation*, vol. 11, no. 5, May 2023.

[7] K. Tanabe, "Machine-Learning-Based Composition Analysis of the Stability of V-Cr-Ti Alloys," *J. Nucl. Eng.*, vol. 4, no. 2, pp. 317–322, Apr. 2023.

[8] J. Datta, D. Datta, and V. Sharma, "Transferable and Robust Machine Learning Model for Predicting Stability of Si Anodes for Multivalent Cation Batteries," *J. Mater. Sci.*, vol. 58, no. 27, pp. 11085–11099, 2023.

[9] C. Hong *et al.*, "Training machine-learning potentials for crystal structure prediction using disordered structures," *Phys. Rev. B*, vol. 102, no. 22, p. 224104, 2020.

[10] A. Ihalage and Y. Hao, "Analogical discovery of disordered perovskite oxides by crystal structure

information hidden in unsupervised material fingerprints," *npj Comput. Mater.*, vol. 7, no. 1, p. 75, 2021.

[11] S. Dahal, D. Adhikari, and S. K. Yadav, "Machine Learning Model to Predict the Formation Energy of Copper-based Ternary Alloys," *J. Inst. Sci. Technol.*, vol. 29, no. 2, pp. 99–105, 2024.

[12] Z. Fan *et al.*, "Carbon-Free Conversion of SiO2 to Si via Ultra-Rapid Alloy Formation: Toward the Sustainable Fabrication of Nanoporous Si for Lithium-Ion Batteries," *ACS Appl Mater Interfaces*, vol. 15, no. 30, pp. 36076–36085, Aug. 2023.

[13] S. Kaiser and A. Khan, "Role of silicon on the tribological performance of Al-based automotive alloys and the effect of used motor oil," *Tribol. - Finnish J. Tribol.*, vol. 39, no. 3--4, pp. 12–20, Dec. 2022.

[14] Y. Liu, "The silicon-based anode in lithium-ion battery," in *Journal of Physics: Conference Series*, Institute of Physics, 2022.

[15] M. K. Majeed *et al.*, "Silicon-based anode materials for lithium batteries: recent progress, new trends, and future perspectives," *Crit. Rev. Solid State Mater. Sci.*, vol. 49, no. 2, pp. 221–253, Mar. 2024.

[16] A. M. Numan-Al-Mobin and A. Smirnova, "Silicon-based lithium-ion battery anodes and their application in solid-state batteries," in *Green Sustainable Process for Chemical and Environmental Engineering and Science*, Elsevier, 2023, pp. 129–169.

[17] T. Fazal, F. Ali, N. S. Hosmane, and Y. Zhu, "Boron compounds for catalytic applications," 2022, pp. 169–199.

[18] G. M. Gorito, M. F. Vieira, and L. M. M. Ribeiro, "The Role of Boron on the Microstructure and Properties of a Ni-Si-B Cast Alloy," p. 53, Jun. 2022.

[19] N. B. Pugacheva and P. A. Polyakov, "The effect of boron on the protective properties of aluminide coatings," *Mater. Proc.*, vol. 8, no. 1, p. 53, 2022.

[20] B. Schurink, W. T. E. van den Beld, R. M. Tiggelaar, R. W. E. van de Kruijs, and F. Bijkerk, "Synthesis and Characterization of Boron Thin Films Using Chemical and Physical Vapor Depositions," *Coatings*, vol. 12, no. 5, p. 685, May 2022.

[21] L. Ward *et al.*, "Matminer: An open source toolkit for materials data mining," *Comput. Mater. Sci.*, vol. 152, pp. 60–69, 2018.

[22] F. Pedregosa, "Scikit-learn: Machine learning in python Fabian," *J. Mach. Learn. Res.*, vol. 12, p. 2825, 2011.

[23] A. Jain *et al.*, "Commentary: The Materials Project: A materials genome approach to accelerating materials innovation," *APL Mater.*, vol. 1, no. 1, p. 11002, 2013.

[24] S. Raschka, "MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack," *J Open Source Softw*, vol. 3, no. 24, p. 638.