Solid State Communications 361 (2023) 115062

Contents lists available at ScienceDirect

Solid State Communications

journal homepage: www.elsevier.com/locate/ssc

Crystal structural prediction of perovskite materials using machine learning: A comparative study

Rojalina Priyadarshini [a], Hillol Joardar [b], Sukant Kishoro Bisoy [a], Tanmaya Badapanda [c,*]

[a] Department of Comp. Sc.& Engg, C.V.Raman Global University, Bhubaneswar, Odisha, India
[b] Department of Mechanical Engineering, C.V.Raman Global University, Bhubaneswar, Odisha, India
[c] Department of Physics, C.V.Raman Global University, Bhubaneswar, Odisha, India

ARTICLE INFO

Communicated by: Ralph Gebauer

Keywords:
Perovskite
Machine learning
Boosting algorithms
K-nearest neighborhood
XGBoost

ABSTRACT

In this study, Machine Learning (ML) techniques have been exploited to classify the crystal structure of ABO_3 perovskite compounds. In the present work, seven different ML algorithms are applied to the experimentally determined crystal structure data. The relevance of the data featured is measured by computing the Chi-Square test and Spearman's correlation matrix. The Z-Score value has been calculated for each attribute to confirm the existence of any outliers in the data. The Synthetic Minority Oversampling (SMOTE) technique is employed to overcome the imbalanced data set. The models' performance is calculated using the stratified k-Fold cross-validation method. Further, to improve the accuracy of the prediction model, the conventional algorithm is supported by boosting algorithm. Comparative model efficiency on prediction of the crystal structure is presented to identify the most suitable model. As per the inferences drawn from the observations, the ensemble model using Xtreme Gradient Boosting (XGBoost) algorithm when applied to the pre-processed and balanced data outperforms the other models.

1. Introduction

ABO_3 perovskite compounds have been discovered over the past decades due to their various industrial applications [1]. Perovskites are one of the most common structural families, and they can be found in a broad range of compounds with diverse properties, applications, and significance [2]. Fig. 1 demonstrates the cubic, orthorhombic, rhombohedral and tetragonal crystal structure of ABX_3 perovskite material [3]. The composition of an ideal perovskite is ABX3, with 'A' and 'B' being two different cations and 'X' being an anion ('X' is considered as oxygen in the present work). In the dodecahedral setting, the 'A' cation is encircled by 12 oxygen ions, while the 'B' cation is octahedrally coordinated by 6 oxygen ions. This material exhibits a diversity of attractive physical performance thatinvolved ferroelectric, dielectric, and piezoelectric behavior [4–7]. The above-mentioned properties of perovskite materials have displayed the potential to be employed in a wide range of applications. The properties of the perovskite compounds mostly depend on the crystal structure [8]. Principally, the crystal arrangements of the perovskite-type compositions, strongly influence the electronic arrangements and put an impact on the energy band gaps and carrier transport [9]. Hence, the research community needs to identify the crystal structure before designing aspirant materials for the next-generation application. The conventionalmeans to synthesize new materials are typically based on the trial and error concept and continuous synthesis and characterization continue until thematerials meet the target. This method needs long-time learning on a limited amount of materials and complex experimental processes, which can be a lengthy and expensive attempt. In addition, the sighting of high-performance materials requires a long span from experimental design to commercialization. It is an immenseconfront to figure out the connection between materials constructions and performance by customary experimental methods.

As experimentally crystal structure identification is time-consuming and costly, an alternative to this problem is quite demanding in the present scenario. In this regard, a few simulation techniques such as Density Functional Theory (DFT) [10], Monte Carlo simulation [11], and molecular dynamics [12] are utilized to investigate the connection between the structural, compositional, and performance of materials at different scales. Nevertheless, most computational schemes only aim at a definite system, directing an awful amount of computation for complex systems. Moreover, computational simulation methods require high computational costs and professional skills. Further, this approach is

[*] Corresponding author.
E-mail address: badapanda.tanmaya@gmail.com (T. Badapanda).

https://doi.org/10.1016/j.ssc.2022.115062
Received 29 October 2022; Received in revised form 22 December 2022; Accepted 23 December 2022
Available online 29 December 2022

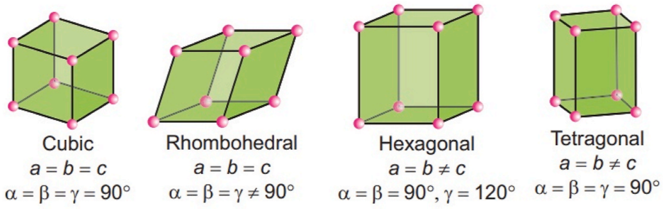
Cubic
$a = b = c$
$\alpha = \beta = \gamma = 90°$
Rhombohedral
$a = b = c$
$\alpha = \beta = \gamma \neq 90°$
Hexagonal
$a = b \neq c$
$\alpha = \beta = 90°, \gamma = 120°$
Tetragonal
$a = b \neq c$
$\alpha = \beta = \gamma = 90°$

Fig. 1. Various perovskite oxide crystal structure of ABX3 compounds.

very challenging when it deals with compounds requiring large unit cells and functional cells much larger than the primitive ones. Machine Learning based approach is found to be one of the suitable processes that cruise through the experimental data to form a statistical framework for the prediction of the stability of the compound [13]. To reduce the huge time required to conduct large-scale screening studies, machine-learning approaches have been demonstrated to efficiently predict many properties of materials. A series of achievements in ML has been made in various kinds of materials and their performance such as superconductors, photovoltaic materials, and high entropy alloys [14–16]. Zhang and Xu [17] predicted lattice parameters for orthorhombic distorted perovskite oxides using the Gaussian process regression (GPR) model and it shows promising prediction power as well. Zhang et al. [18] have reviewed recent progress in the machine learning studies of halide perovskite materials, including the prediction and understanding of lead-free and stable halide perovskite materials. Balachand et al. [19] constructed classification ML models with a 90% average accuracy and predicted 235 structurally stable ABO3 perovskites among a list of 625 new ABO3 compounds. Earlier some metaheuristic optimization algorithms like Particle Swarm Optimization (PSO), genetic algorithm, and Ant Colony Optimization (ACO) along with some evolutionary algorithms are being used for stable crystal structure prediction [20]. The popular software packages like USPEX, COPEX are inspired by the evolutionary computing [21,22]. The idea behind these algorithms is to evaluate among multiple possible composition and to come up with a composition having minimum energy level. The CALYPSO is a software package used to predict the crystal structures of the materials from the chemical compositions along with some external condition such as pressure [23]. It is based on the particle swarm optimization algorithm. Its basic intention is to obtain an energetically stable structure of the material. These methods work well for the simple binary systems, but for the ternary complex systems, the number of possible compositions may increase exponentially as per the increase in the quantity of atoms present in a single cell. Whereas the ML algorithms work well for a larger number of data and the models can better generalize and can find the hidden patterns in the data unless the data is curated properly. Choosing the right kind of parameters and data cleaning is a challenging task for the ML algorithms. Most of the works reported earlier are mostly focused on one category of classification of prediction of crystal structure with limited input parameters for a limited size of the database. In the present work, we focus on the prediction of classifying ABO_3 compounds in cubic, tetragonal, orthorhombic, and rhombohedral systems with a large number of input parameters including vacancy, ionic radii, electronegativities, and oxidation states. The main task of this paper is to study the available ABO3 perovskite data and to exhibit the inherent characteristics of the data. To do so, the correlation among each attribute concerning the class attribute is statistically calculated. Which gives a clear distinction about how much each of the attributes is contributing to identifying the class attribute. The obtained data is found to contain non-uniform distribution among the four classes of the available data. This is a generic problem for the accurate determination of a class. Majors are taken to make the data balanced which will contain a nearly equal number of samples in each class. Finally to build Machine Learning (ML) models to predict the crystal structure with a good degree of precision from the available data on crystal structure classification of ABO3 solids and to present a comparative result of the model to obtain the optimized model. The major contribution of this work is enlisted as follows.

I. An array of ML models such as K-Nearest Neighborhood (KNN), Naïve Bayes, Support Vector Machine (SVM), Multilayer feed-forward Artificial Neural Network (ANN), and a couple of Boosting algorithms such as Light Gradient Boosting method and Extreme Gradient Boosting (XGBoost) are utilized which are aiming to categorize ABO3 perovskite compounds to four different structural classes
II. A rigorous feature selection mechanism is carried over the collected data to get rid of irrelevant features. In this regard, the Chi-Square test is performed among the class and categorical attributes. Pearson's correlation matrix is calculated to check the redundancy of the attributes.
III. To improve the performance of the prediction model SMOTE is used which helps to handle the imbalanced data problem along with the stratified K-Fold cross-validation process which in turn helps to increase the overall performance and fight against the overfitting problem which is a problem in the user data.

2. Methodology

To accomplish the enlisted contribution a working model has been proposed which consists of the following steps. The proposed model has been validated by conducting a set of experiments on perovskite data collected from Ref. [24]. The flow diagram of the proposed model to predict the crystal structure is represented in Fig. 2. The function of each phase is explained in the below section.

2.1. *Dataset collection*

In the present work, ABO3 perovskite-type oxides are being considered and different analysis has been carried out with the respective data and their attributes [24]. A complete selected feature list to do the task is mentioned in Table 1. The dataset has 5329 samples in total having 15 features as depicted in Table 1. The table includes the attributes that have been chosen for the task, as well as information about any missing values in the columns and the presence of outliers. The data points whose standard deviations are far away from the mean are considered outliers. The outliers are obtained by determining the Z-score of each column by using equation (1). The data are tested for extreme positive and negative values according to the calculated Z-score of each column. The extreme value indicates the presence of outliers and the presence of outliers is depicted in Table 1. It is perceived that the 'Bond length A-O′ and 'Bond length B–O′ are found to contain zero and are further filled with mean values of the rest data of the column (see Table 2).

$$Z=\frac{x-\mu}{\sigma}Z=\frac{x-\mu}{\sigma} \tag{1}$$

where X is the observed column, μ is the mean of the column and σ is the standard deviation of the same column.

2.2. *Data pre-processing*

The performance of a machine learning model depends heavily on the pre-processing and cleaning of the data. The considered dataset for this experiment contains some redundant values, null values, and missing values. The primary goal of the current phase is to get rid of all these parameters which may affect the overall efficiency of the model. A set of tasks is being done to treat the missing values, deal with the categorical attributes, and handle the scaling difference among the attributes.

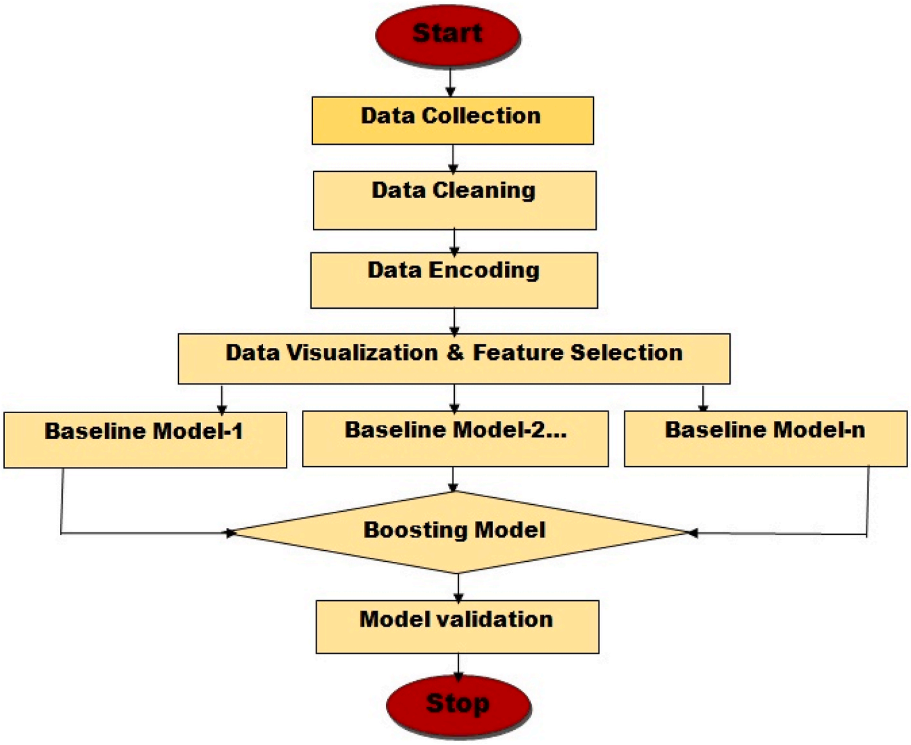

Fig. 2. Flow diagram of the proposed ensemble model using boosting methods to predict the crystal structure.

Table 1
Characteristics of data used for the proposed work.

Features	Type of data	Presence of Missing value	Presence of Outliers
Compound	Categorical	No	No
A	Categorical	No	No
B	Categorical	No	No
In literature	Boolean	No	No
Valence (A)	Mixed	No	No
Valence (B)	Mixed	No	No
Radius A	Numerical	No	No
Radius B	Numerical	No	No
Electro negativity of A	Numerical	No	No
Electro negativity of B +	Numerical	No	No
Bond Length of A-O	Numerical	No, some have ‘0’	Yes
Bond Length of B–O	Numerical	No, some have ‘0’	Yes
Electronegativity difference with radius	Numerical	No	No
Goldschmidt tolerance factor	Numerical	Yes	Yes
New tolerance factor	Numerical	No	No
Octahedral Factor	Numerical	No	No
Lowest Distortion	Categorical	No	No

Table 2
Shows the Chi-Square test values for the categorical values.

Columns	Chi-quare P-value	Rounded P-value	Null Hypothesis Observation
Compound	1.17637185107542258e-199	0	Rejected
V(A)	2.7058568585255008e-48	0	Rejected
V(B)	2.7187923908884558e-38	0	Rejected
In literature	1.2838904300832779e-117	0	Rejected

2.3. Handling of missing value

The prime objective of the present step is to handle the null values as well as the values of the attributes captured as ‘0’. By keeping an insight into the data, it has been observed that the used data is having some null values as ‘0’ within it. Therefore, the column values containing ‘0’ are replaced by the mean of remaining values that are not ‘null’. Also, the attribute named ‘New Tolerance Factor’ is obtained by possessing some random and irrelevant values like ’-’. Hence, the symbol ’-’ is replaced by ‘0’ as it indicates the absence of octahedral values for certain instances.

2.4. Dealing with categorical values

The present dataset contains a couple of feature attributes that contain categorical values including the class attribute. The three ‘Valence’ attributes have categorical data which are converted into numerical data. The target class is transformed using the Python “Label Encoder” package, which modifies all five of the labels that are present in the dataset between 0 and 1. One hot encoding is being employed to convert the other categorical attributes to numerical ones.

2.5. Normalization

The large difference in scale of values of different features greatly impacts the performance of the model. Normalization is a process to bring down all the attributes within a closer range. It helps to do the training in the right and stable way. In the present work, Min-Max normalization is being taken over the entire dataset so that all the feature values are kept within a closer range.

2.6. Feature selection

The selection of features from a dataset is a very crucial task when it comes to training and testing models. The feature decides the adequacy of the trained model and its performance. The correct choice of features thus becomes crucial before going for the algorithm selection process.

The embedded method is being used to do the feature selection from the collected dataset. The Chi-square Test for independence is carried out to know the level of interdependency among the class and categorical attributes. The formula for the Chi-Square test of independence is presented in equation (2):

$$x^2 = \sum \frac{(O_i - E_i)^2}{E_i} x^2 = \sum \frac{(O_i - E_i)^2}{E_i} \tag{2}$$

where χ^2 is the chi-squared statistics and Oi and Ei are the Observed and Expected value of frequency of two of the categorical attributes and 'i' is the total number of cells in the contingency table. In this work, the Chi-Square test is done on the class attribute and all the categorical attributes to check the dependency between them. The heat map of the contingency tables is illustrated in Figs. 3–5. In the performed Chi-Square test, the 'p-Value' is coming closer to '0'. It indicates that the null hypothesis is rejected. According to the final observation, the categorical values are also dependent on the class attribute, thus, they play a significant role during model training. Table −2 illustrates the P-values obtained during the Chi-Square test for the categorical attributes.

The correlation value is calculated using Spearman's correlation coefficient for all the attributes and that is shown in Fig. 6. The formula for Spearman's correlation is represented in Equation (3). This correlation provides a true correlation result because it does not make any assumptions about the distribution of the data and performs best for numerical values. The heat map graph of the same represents a clear correlation among all the attributes. Since the used dataset now contains 675 items and the category class distribution is significantly unbalanced, the cell with a high positive value denotes a poor correlation, which suggests that the proportion of samples in each class is not the same. It can be removed from the feature set while training. In the present case the feature 'Radius B-r(BVI)(A)'is found to be the highest association with 'Octahedral factor-μ', and it becomes redundant to be considered for training purposes. Therefore, during the training phase, only the 'Octahedral factor-μ' is considered.

$$\rho = 1 - \frac{6\sum di^2}{n(n^2-1)} \rho = 1 - \frac{6\sum di^2}{n(n^2-1)} \tag{3}$$

The distribution of data samples of the entire data set is shown in Fig. 7. It is seen from Fig. 7 that the number of samples present in each class is not of the same proportion for their data imbalances. The Synthetic Minority Oversampling Technique (SMOTE) has been employed [25] to get uniform distribution of several samples. SMOTE is an over-sampling technique that helps to synthesize the number of instances in minority classes. It not only duplicates the samples in the minority class, rather it works like a data augmentation technique that produces new instances resembling closer to the original instances in the minority class and also is a suitable way to deal with tabular kind of data [26].

3. Models used for structure classification

Despite the existence of numerous machine learning (ML) techniques, the "No free lunch theorem" of approximation and optimization states that no single model can be relied upon to perform effectively across all the problems [27]. Four kinds of perovskite compounds are included in the dataset taken into account for the experiment. Therefore, all of the proposed models are chosen to perform multiple class classification, allowing them to divide the ABO3 perovskites compounds into four groups, namely 1) Cubic 2) Tetragonal 3) Orthorhombic 4) Rhombohedral. The dataset has undergone the essential cleaning and pre-processing process. Chi-Square test experiments are used to choose the feature, and Spearman's correlation is used to compute the numerical correlation. Different models are evaluated on the data to achieve the greatest performance, and the experiment is started with the KNN algorithm. The value of 'K' plays a crucial role in deciding the performance of the KNN model. The error rate is therefore compared to the different values of "K" in the experiment. The optimal value of 'K' is settled at 6 with a maximum accuracy rate of 79%. The graph is drawn

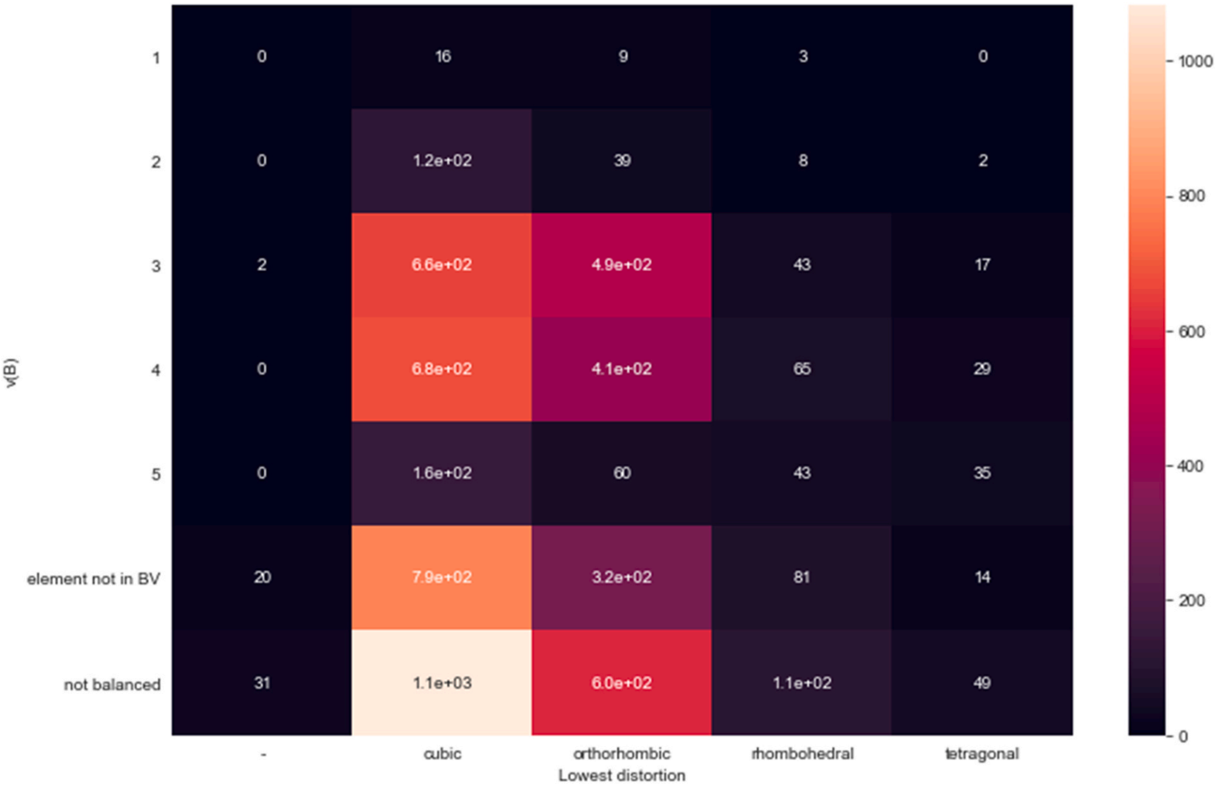

Fig. 3. Heatmap representation of contingency tables of class and column valance of B.

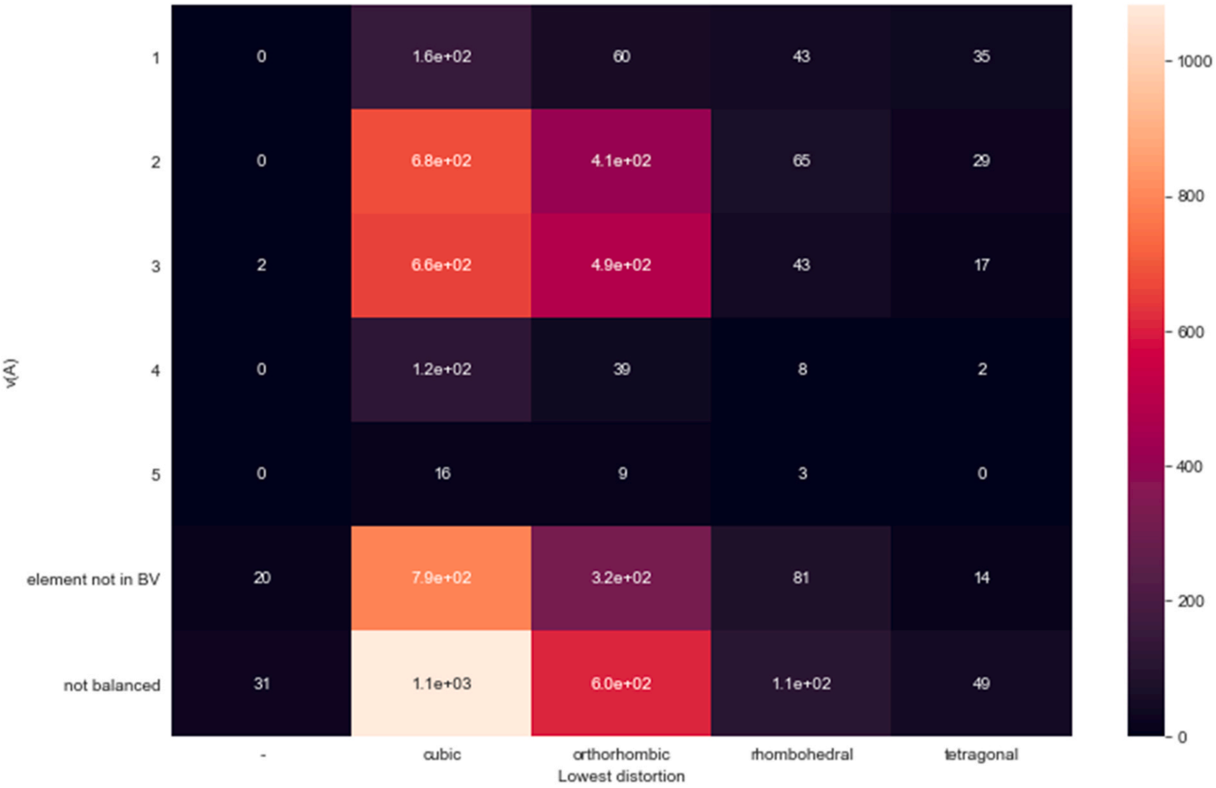

Fig. 4. Heatmap representation of contingency tables of class and column valance of A.

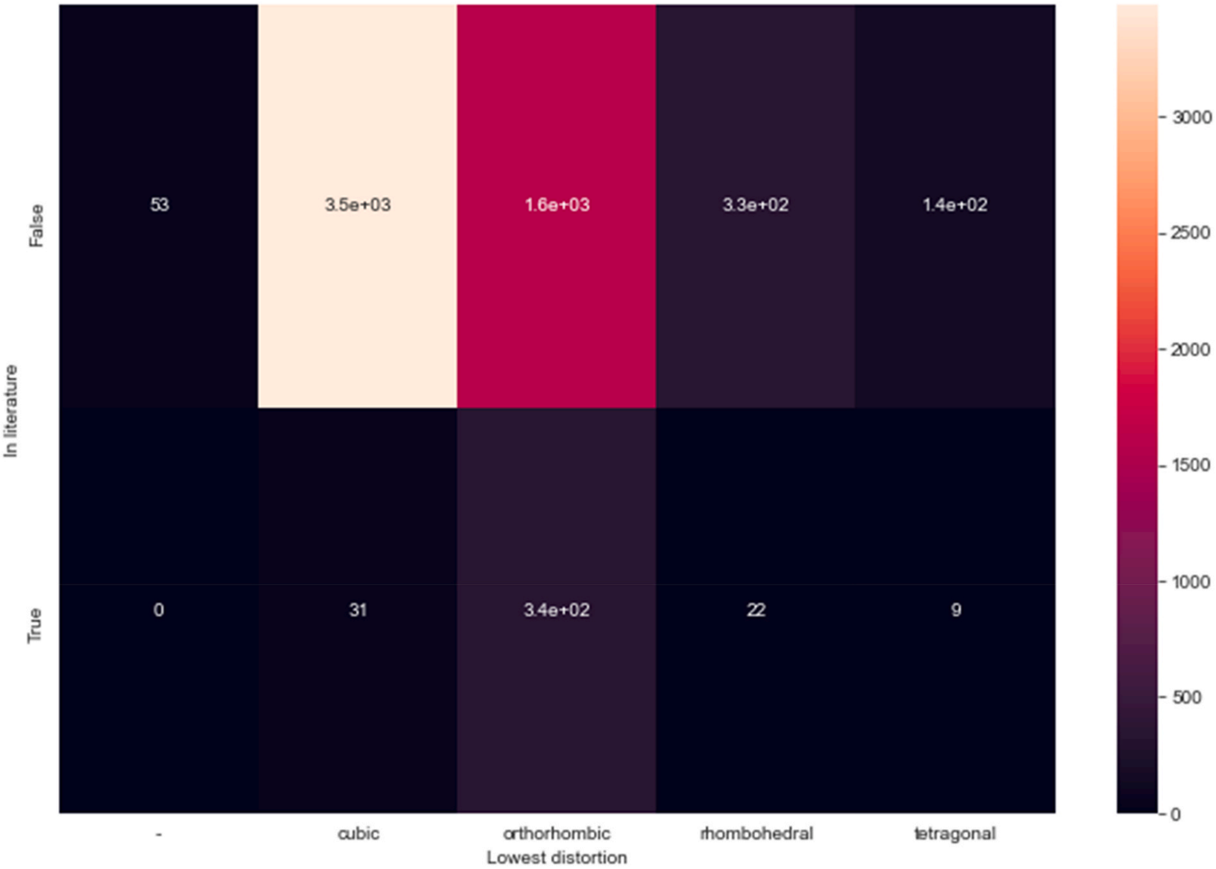

Fig. 5. Heatmap representation of contingency tables presenting Pervoskitee reported in litearature

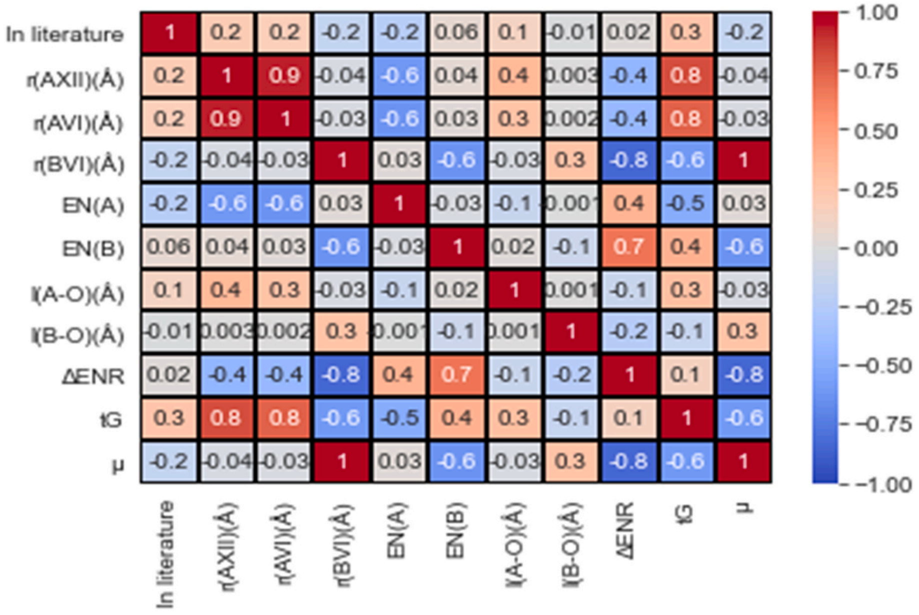

Fig. 6. Feature correlation matrix for formability prediction.

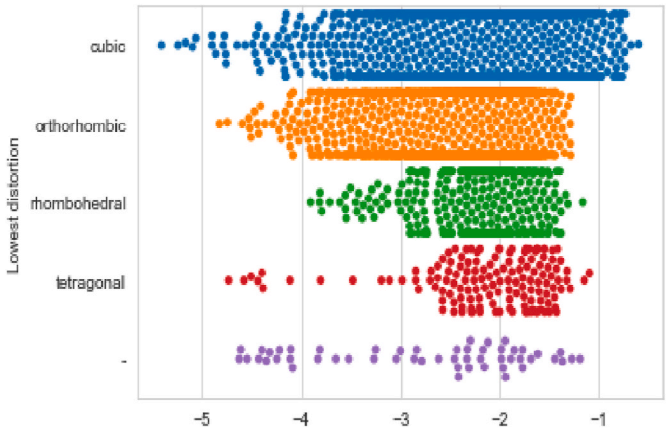

Fig. 7. Distribution of data samples.

using various "K" values and the calculated error rate, as shown in Fig. 8. Then the Naïve Bayes algorithm is executed over the dataset and that gives around 71% of accuracy. The SVM classifier is chosen with the Radial basis kernel and ended with 66% accuracy. The ANN architecture chosen in the work is a multilayer feed-forward algorithm that has three hidden layers with a sigmoidal activation function. The gradient descent optimization technique is used as a cost function and the learning rate parameter is chosen as 0.03. The ANN gives an accuracy of 71.46%.

A decision tree is a subsequent classifier utilized for the task. A decision tree is a powerful tool that could be employed for several tasks such as classification and regression. It is a non-parametric algorithm. In the present experiment, the split of the tree is determined by computing information gain using the ID3 algorithm. In comparison to ANN and Naive Bayes algorithms, the decision tree that has been used performs very well. By doing a loop across the tree and evaluating the accuracy at each value of the depth taken, the depth of the tree is ultimately decided to be 6. The dataset is then tested with a few ensemble algorithms after being subjected to testing with all possible classifier types. Light GBM and XGBoost are the two boosting techniques that have been selected to achieve this. Both have performed better than others but XGBoost has outperformed all the classifiers in terms of accuracy. XGBoost is an ensemble technique based on the decision tree-based machine learning model. It is suitable for small to medium-sized structured data and even performed well with a mixture of numerical and categorical data. These factors make XGBoost the ideal choice for the data used in this work.

4. Experimental results and analysis

The entire experiment is carried out by Python version 3 using the SciKit Learn library. In this case, the Rate of misclassification or accuracy is taken as evaluation parameters. The total number of predicted positive rates or 'Precision' and True positive rate or 'Recall'. To improve performance and prevent the overfitting issue and increase the reliability of the model stratified cross-validation is utilized to divide both the training and test data into 10-fold groups. The model assessment is therefore performed on each fold of test data to obtain the final set of results, and the results that are provided of the average of all folds.

Fig. 8 shows the stratified k-fold cross-validation by using test data set. Cross-validation is a useful technique that can aid the supervised machine learning model's ability to generalize well as compared to the holding method [28]. A K-fold cross-validation approach randomly divides the test dataset into k folds, one of which is utilized for testing and the other for training. This procedure is repeated for each value of 'k'. Compared with K-fold and stratified K-Fold cross-validation schemes, the latter helps to get rid of data imbalance problems occurring during training and test split. Because in stratified K-fold cross-validation, it is ensured that each fold has a balanced amount of samples from each class. The value of 'K' is chosen as 10 hereafter experimenting with different K-values starting with 2–10. The effect of the rate of change in error concerning the random number of folds on test data is depicted in Fig. 9. Before choosing the Boosting methods, the experiment is carried out by taking into account some of the well-known classifiers, such as KNN, decision tree, SVM, Nave Bayes, Multilayer ANN, and Light GBM. The boosting algorithms are designed to increase accuracy, and the outcomes of the experiments have supported this claim. The performance of all the algorithms is presented in Table 3. The XGBoost offers superior accuracy, precision, and recall compared to other options, which supports the author's selection.

5. Conclusion

Various ML techniques such as KNN, Naïve Bayes, SVM, Multilayer feed-forward ANN, Light GBM, and XGB are used for the prediction of

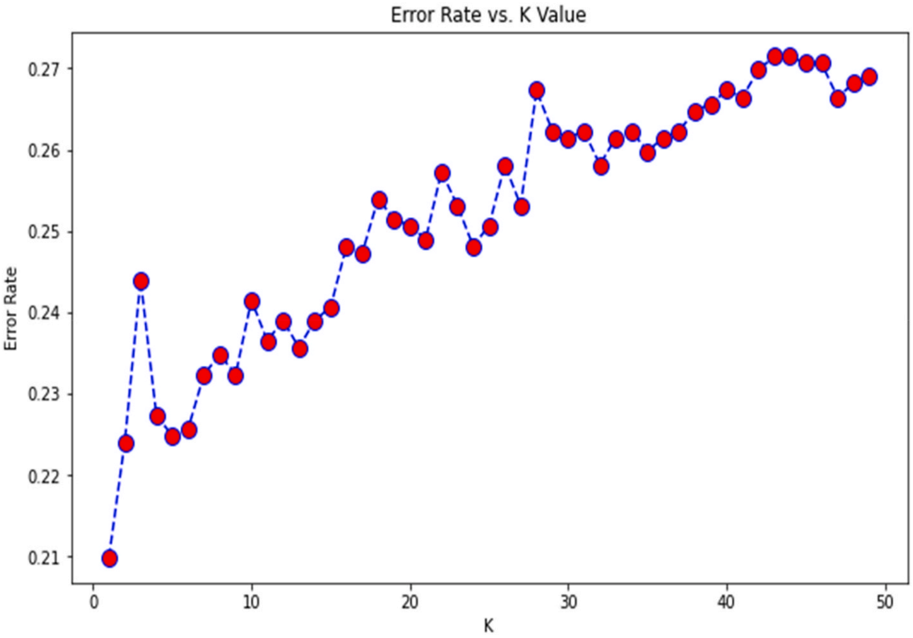

Fig. 8. Variation of K values with the error rate in the KNN algorithm.

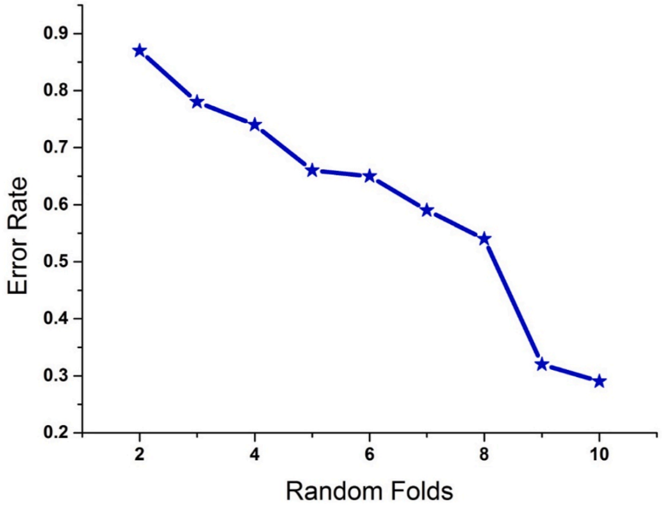

Fig. 9. Variation of error rate with random fold taken on the test data.

Table 3
Performance comparison of Baseline Models.

Sl. No.	Algorithm	Accuracy%	Precision	Recall	F1-score
1	KNN	74.11	74.56	76.53	75.11
2	Decision Tree	76.14	74.20	75.56	74.21
3	SVM	65.17	64.17	66.11	63.98
4	NB Classifier	70.76	64.65	70.89	65.67
5	ANN	71.46	64.43	71.67	66.43
6	Light GBM	80.79	80.17	80.32	79.10
7	XGB	83.67	82.10	82.12	82.11

perovskite crystal structure. Out of a total of 5329numbers of available perovskite, the experimentation is done on 675 compounds, which are classified successfully into four categories. A range of activities has been done to understand the underlying behavior of the data and to get detailed insight. The experimental data were pre-processed to eradicate the duplicate values and to handle the categorical attributes. The correlation among the data was measured using Spearman's correlation matrix. Both numerical and categorical types of the feature have been studied against the class attributes. The Chi-square test has been used for categorical characteristics because data pre-processing and feature selection are two key criteria that have a significant impact on a model's performance. The imbalanced nature of the data is handled using SMOTE. The experiment has carried out on the processed data by optimizing various adjustable parameters of multiple models to produce a classifier that predicts the structure of a perovskite crystal system automatically. The overall performance of the models was measured using the K-Fold cross-validation method. The XGBoost algorithms have produced an accuracy of 84% among all the examined models, which is an acceptable rate to utilize and consequently offers a quick and affordable technique to determine the crystal structure. In the next work, some more relevant features are to be fetched and their behavior will be studied as far as their structure is concerned. Experiments will be performed in this direction, which will add value and impact to automatic ML-based structure categorization techniques.

Funding

No funding was received for conducting this study.

Code availability

Not applicable.

Author contribution

Conceptualization, Tanmaya Badapanda; Data curation, Sukanta Bisoyi; Formal analysis, Hillol Joardar.; Methodology& Implementation, Rojalina Priyadarshini.

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.

Data availability

Data will be made available on request.

References

[1] EMRAssirey, Perovskite Synthesis, Properties and Their Related Biochemical and Industrial Application, vol. 27, 2019, pp. 817–829, https://doi.org/10.1016/j.jsps.2019.05.003.

[2] R.F. Service, Materials scientists look to a data-intensive future, Science 335 (2012) 1434, https://doi.org/10.1126/science.335.6075.1434.

[3] Cristina Artini, Crystal chemistry, stability and properties of interlanthanide perovskites: a review, J. Eur. Ceram. Soc. 37 (2) (2017-02-01) 427–440, https://doi.org/10.1016/j.jeurceramsoc.2016.08.041.

[4] O. Muller, R. andRoy, The Major Ternary Structural Families, Springer, NewYork, 1974.

[5] N.A. Hill, J. Phys. Chem. B 104 (2000) 6694.

[6] J.F. Scott, Ferroelectr. Rev. 1 (1998) 1.

[7] A.J. Millis, Nature 392 (1998) 147.

[8] T. Han, S. Ma, X. Xu, P. Cao, W. Liu, X. Xu, S. Pei, Electrospinning synthesis, crystal structure, and ethylene glycol sensing properties of orthorhombic SmBO3 (B=Fe, Co) perovskites, J. Alloys Compd. 876 (2021) 160–211, https://doi.org/10.1016/j.jallcom.2021.160211.

[9] V. Kažukauskas, V. Janonis, V. Vertelis, Energy band-gap inhomogenities and defect states affecting carrier transport at low temperatures in Thallium Bromide, Opt. Mater. 118 (2021) 111–259, https://doi.org/10.1016/j.optmat.2021.111259.

[10] A.R. West, Basic Solid State Chemistry, second ed., John Wiley & Sons, New York, 1999.

[11] P. Hohenberg, W. Kohn, Inhomogeneous electron gas, Phys. Rev. 136 (1964) B864–B871.

[12] A. Hussain, et al., Monte Carlo simulation study of electron yields from compound semiconductor materials, J. Appl. Phys. 128 (2020), 015305.

[13] P.V. Balachandran, B. Kowalski, A. Sehirlioglu, T. Lookman, Experimental search for high-temperature ferroelectric perovskites guided by two-step machine learning, Nat. Commun. 9 (2018) 1668.

[14] D. Dai, et al., Using machine learning and feature engineering to characterize limited material datasets of high-entropy alloys, Comput. Mater. Sci. 175 (2020), 109618.

[15] A. Mannodi-Kanakkithodi, G. Pilania, T.D. Huan, T. Lookman, R. Ramprasad, Machine learning strategy for accelerated design of polymer dielectrics, Sci. Rep. 6 (2016), 20952, https://doi.org/10.1038/srep20952.

[16] W. Sun, et al., Machine learning-assisted molecular design and efficiency prediction for high-performance organic photovoltaic materials, Sci. Adv. 5 (2019), eaay4275.

[17] V. Stanev, et al., Machine learning modeling of superconducting critical temperature, NPJ Comput. Mater. 4 (2018) 29.

[18] Yun Zhang, Xiaojie Xu, Predicting lattice parameters for orthorhombic distorted-perovskite oxides via machine learning, Solid State Sci. 113 (March 2021), https://doi.org/10.1016/j.solidstatesciences.2021.106541.

[19] Lei Zhang, He Mu, Shaofeng Shao, Machine learning for halide perovskite materials, Nano Energy 78 (December 2020), https://doi.org/10.1016/j.nanoen.2020.105380.

[20] P.V. Balachandran, D. Xue, J. Theiler, J. Hogden, T. andLookman, Adaptive strategies for materials design using uncertainties, Sci. Rep. 6 (2016), 19660, https://doi.org/10.1038/srep19660.

[21] Daniel Chu, Zomaya Albert, Parallel ant colony optimization for 3D protein structure prediction using the HP lattice model, in: Parallel Evolutionary Computations, Springer, Berlin, Heidelberg, 2006, pp. 177–198.

[22] Yanchao Wang, Jian Lv, Li Zhu, Yanming Ma, CALYPSO: a method for crystal structure prediction, Comput. Phys. Commun. 183 (10) (2012) 2063–2070.

[23] Andriy O. Lyakhov, R.Oganov Artem, Harold T. Stokes, Qiang Zhu, New developments in evolutionary structure prediction algorithm USPEX, Comput. Phys. Commun. 184 (4) (2013) 1172–1182.

[24] Xiangyang Liu, Haiyang Niu, Artem R. Oganov, COPEX: co-evolutionary crystal structure prediction algorithm for complex systems, npj Computational Materials 7 (1) (2021) 1–11.

[25] Santosh Behara, Taher Poonawala, Tiju Thomas, Crystal structure classification in ABO3 perovskites via machine learning, Comput. Mater. Sci. 188 (2021), 110191, https://doi.org/10.1016/j.commatsci.2020.110191.

[26] A. Gosain, S. Sardana, Handling class imbalance problem using oversampling techniques: A review, in: 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Udupi, India, 2017, pp. 79–85, https://doi.org/10.1109/ICACCI.2017.8125820.

[27] Alberto Fernández, Salvador Garcia, Francisco Herrera, Nitesh V. Chawla, SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary, J. Artif. Intell. Res. 61 (2018) 863–905, https://doi.org/10.1613/jair.1.11192.

[28] Stavros P. Adam, N. Stamatios-Aggelos, Panos M. Pardalos Alexandropoulos, Michael N. Vrahatis, No free lunch theorem: a review, Approximat. Optimizat. (2019) 57–82, https://doi.org/10.1007/978-3-030-12767-1_5.