Computational Materials Science 210 (2022) 111476

Contents lists available at ScienceDirect

Computational Materials Science

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

Full Length Article

# Discovery of direct band gap perovskites for light harvesting by using machine learning

Smarak Rath [a], Sudha Priyanga G. [b], Nagappan N. [a], Tiju Thomas [a,*]

[a] Department of Metallurgical and Materials Engineering, Indian Institute of Technology Madras, Chennai, Tamil Nadu 600036, India
[b] Department of Physics, Mepco Schlenk Engineering College, Sivakasi, Tamil Nadu 626005, India

ARTICLE INFO

*Keywords:*
Perovskite
Machine learning
Nature of band gap
XGBOOST
Matminer

ABSTRACT

An approach that would allow quick determination of compositions that are most likely to be direct band gap materials would significantly accelerate research on light-harvesting materials. Inorganic perovskites are attractive for this purpose since they afford compositional flexibility, while also offering stability. Here, $ABX_3$ inorganic perovskites (A and B are cations and X is an anion) are classified into direct band gap and indirect band gap materials by using the XGBOOST (eXtreme Gradient BOOST) classifier. We use a dataset containing 1528 $ABX_3$ compounds (X = O, F, Cl, Br, I, S, Se, Te, N, or P) along with information on the nature of their band gap (direct or indirect). All the data is taken from the Materials Project database. Descriptors for these materials are generated using the Matminer python package. Ten-fold cross-validation with the XGBOOST classifier is used on the dataset and the average accuracy is found to be 72.8%. To generate a confusion matrix, the dataset is once again split into a training set and a testing set after cross-validation. Subsequently, the confusion matrix is generated for that particular test set. It is found that the precision for the prediction of direct band gap materials is 81% i.e., 81% of the materials predicted to be direct band gap materials are actually direct band gap materials. Thus, machine learning can be an effective tool for discovering novel direct band gap perovskites. Finally, SHAP (SHapley Additive exPlanations) analysis is performed to determine the most important descriptors. One key insight gained from the SHAP analysis is that the absence of transition metals and elements belonging to groups IIIA to VIIIA with atomic number greater than 20 increases the probability of the perovskite having a direct band gap.

## 1. Introduction

The fundamental band gap of a material is the most important property for light-harvesting applications. Such materials find use in photocatalytic and photovoltaic applications [1,2]. For manufacturing solar cells with high efficiencies, materials with low band gaps need to be used. Lower is the band gap of the absorber material, higher is the percentage of incident photons it can absorb. It is well known that the optimal band gap of direct gap materials used for making the absorber layer of a solar cell is between 1.1 eV and 1.4 eV [3]. Materials with band gaps higher than 3.1 eV are almost entirely unfit for solar cell applications. Such materials can utilise only the ultraviolet range which contains less than 4% of the energy of the solar spectrum [4]. Currently, the market is dominated by silicon solar cells as Si has a fundamental band gap of 1.1 eV [5] and it is cheap because of its abundance. Some other materials which are used for making the absorber layer of solar cells include CdTe and Copper Indium Gallium Selenide (CIGS) [6].

Among low band gap materials, direct band gap materials are preferable to indirect band gap materials because the former have much higher absorption coefficients. In indirect band gap materials, the valence band maximum and conduction band minimum are located at different values of wave vector k. Hence, the excitation of an electron from the valence band to the conduction band in such materials is possible only when there is a simultaneous interaction between an electron, photon, and phonon. The probability of such an interaction is low and this is the reason behind the lower absorption coefficients of indirect band gap materials [7].

It is very common for materials scientists to perform Density Functional Theory (DFT) simulations to generate the electronic band structure diagrams of various materials. Once the electronic band structure diagram of a material is generated, it can be ascertained whether the material has a direct or an indirect band gap. But DFT simulations can be

* Corresponding author.
*E-mail address:* tt332@cornell.edu (T. Thomas).

https://doi.org/10.1016/j.commatsci.2022.111476
Received 31 January 2022; Received in revised form 14 April 2022; Accepted 23 April 2022
Available online 2 May 2022

extremely time-consuming, especially for a material whose unit cell contains a large number of atoms. The time taken by a DFT simulation scales in the order of $N^3$, where N is the number of atoms in the unit cell of the material [8].

In this work, we have made an effort to overcome the problems posed by DFT by using Machine Learning (ML) for the prediction of the nature of band gap (direct or indirect) of $ABX_3$ perovskites. The calculation of properties using DFT may take many hours or even days and months (if the number of atoms in the unit cell of the material is very high). But the prediction of material properties using machine learning can be accomplished in a matter of minutes [9]. Ever since the launch of the Materials Genome Initiative in 2011 [10], materials scientists have successfully used machine learning to predict various properties of materials like ionic conductivity [11], glass transition temperature [12,13], electron affinity [14], vacancy migration energy [15], etc. Many researchers have used machine learning to study materials for optoelectronic devices [16–18]. Machine learning has also been used to predict various properties of perovskites. Zhang et al. used Gaussian Process Regression (GPR) to predict the maximum magnetic entropy change in $ABO_3$ perovskites [19]. Kim et al. used Kernel Ridge Regression (KRR), Random Forest (RF) and Least Absolute Shrinkage and Selection Operator (LASSO) to predict the dielectric breakdown strength of $ABX_3$ perovskites [20]. Xu et al. used Support Vector Machine (SVM) to predict the ionic conductivity of $ABO_3$ perovskites [21]. Takahashi et al. used Random Forest classification algorithm to classify the band gap of perovskites into one of the following categories: 0 – 1.7 eV, 1.7 – 3.0 eV or greater than 3.0 eV [22]. Given the recent reports on the promise of inorganic perovskites for photocatalysis and photovoltaics [23–25], it is reasonable to focus on developing a tool that allows pursuits of direct band gap inorganic perovskites. To the best of our knowledge, there is no report on the use of machine learning for the prediction of the nature of band gap of inorganic perovskites. In the next paragraph, we highlight the reasons for which the investigation of perovskites is a worthwhile endeavor.

Perovskite materials have received a lot of attention in recent years because of their potential for photovoltaic applications. Research on Perovskite Solar Cells (PSCs) gained momentum from 2009. For traditional photovoltaic technologies, the time taken for the technology to become sufficiently evolved has been 15–40 years [5]. PSCs, on the other hand, achieved maturity within 10 years. As of now, the efficiency of perovskite solar cells has already crossed 24% [26]. It may be noted however that this has been the case for organic perovskites. The prospects for inorganic perovskites remain rather underexplored.

An attractive attribute of perovskites is that they exhibit low exciton binding energies. Among perovskites, organic halide perovskites have been studied extensively. In organic halide perovskites, the A-site is occupied by an organic cation (usually $CH_3NH_3^+$), and the X-site is occupied by a halide ion ($Cl^-$, $Br^-$, $F^-$ or $I^-$). Organic halide perovskites have many impressive characteristics like low exciton binding energies, high absorption coefficients, high open circuit voltage and the absence of deep trap states [3]. However, organic perovskites are prone to environmental degradation. The longest lifetime observed for a perovskite solar cell is one year while the lifetime of silicon solar cells is around 25 years [5]. The commercialisation of perovskite solar cells thus depends on increasing their lifetime and making it comparable to that of silicon solar cells.

Inorganic perovskites, unlike organic perovskites, are robust and do not suffer from environmental degradation. Thus, inorganic perovskites can, in principle be used to manufacture solar cells with long lifetimes. The large compositional space that inorganic perovskites span implies that rapid approaches to identify inorganic perovskites with high potential for energy harvesting is a direction worth pursuing. As discussed earlier, the search for direct band gap semiconductors is a crucial activity because materials scientists aim to keep improving the efficiency of solar cells. This work aims precisely to look into this problem i.e., prediction of the nature of band gap of inorganic perovskites using machine learning.

## 2. Computational details

All the data used for this work is taken from the Materials Project database. We use a dataset containing 1528 rows. Each row contains the chemical formula of the $ABX_3$ compound, its Materials Project ID, its space group, and the nature of band gap. The nature of band gap is represented by 0 if the material has an indirect band gap or 1 if the material has a direct band gap. Our ultimate aim is to predict the nature of band gap using machine learning. Our dataset contains multiple polymorphs of some compounds. These polymorphs have the same chemical formula but can be differentiated by their space group and Materials Project ID.

All the work is carried out using python programming in the Jupyter notebook environment. The python packages which we use for our work are Matminer [27], Pymatgen [28], Scikit-Learn [29], XGBOOST [30], SHAP [31], Pandas [32], and Numpy [33]. All the descriptors/features are generated using the Matminer python package. Matminer is an open-source python package that can be used for multiple objectives by materials scientists. It contains 47 classes of featurisation which can generate thousands of descriptors. It can also be used to extract data from popular materials databases like Materials Project [34], Citrination [35], Materials Data Facility [36], and Materials Platform for Data Science [37].

First, we generate pymatgen structure objects for all the materials in our dataset by using MPRester which is the Application Programming Interface (API) used for extracting data from the Materials Project database. MPRester is a part of the Pymatgen python package. Pymatgen structure objects contain information about the unit cell parameters and atomic positions within the unit cell. In order to create crystallographic descriptors with Matminer, the creation of pymatgen structure objects is a must.

After the creation of pymatgen structure objects, crystallographic descriptors are generated by using the 'Structural Heterogeneity' module [38] of Matminer. The generation of crystallographic descriptors is necessary to differentiate between different polymorphs of the same composition. A total of nine crystallographic descriptors are generated which have been listed as follows -mean absolute deviation in relative bond length, maximum relative bond length, minimum relative bond length, minimum neighbor distance variation, maximum neighbor distance variation, range of neighbor distance variation, mean neighbor distance variation, average deviation in neighbor distance variation and mean absolute deviation in relative cell size. Subsequently, 120 other descriptors are generated using the 'Meredig' module [39] of Matminer. These descriptors contain information like atomic fractions of the constituent elements, mean atomic number, mean atomic weight, mean atomic radius, range of atomic radius, mean electronegativity, range of electronegativity, average number of s valence electrons, average number of p valence electrons, average number of d valence electrons, average number of f valence electrons, etc. For the convenience of the readers, the complete list of descriptors used for this work has been provided in Supplementary Information file.

After the generation of descriptors/features, the data in the descriptor columns is standardized using the 'Standard Scaler' module of Scikit-Learn. Subsequently, a correlation matrix is generated, and the highly correlated features are removed. The presence of redundant features can be a serious obstacle if we want to achieve high accuracy [40]. The correlation matrix is generated by calculating Pearson correlation coefficients. The Pearson correlation coefficient gives us the linear relationship between two features. Let us consider that there are two features x and y in a dataset. When the covariance of x and y is divided by the product of their standard deviations, we get the Pearson's correlation coefficient[41] which is denoted by the symbol 'r'. The formula for calculating Pearson's correlation coefficient is:

**Table 1**
List of hyperparameters optimised for each algorithm.

| Algorithm | List of hyperparameters optimised |
|---|---|
| XGBOOST | max_depth, learning_rate, subsample, colsample_bytree, colsample_bylevel, n_estimators |
| Random Forest | n_estimators, max_depth, max_features, min_samples_leaf, min_samples_split, bootstrap |
| Logistic Regression | max_iter, dual |
| Support Vector Classification | gamma, C |
| Artificial Neural Network (ANN) | solver, learning_rate, hidden_layer_sizes, alpha, activation |
| Decision Tree | min_samples_leaf, max_depth, criterion |
| Gradient Boosting Classification | n_estimators, max_depth, learning_rate |

**Table 2**
Values of optimised hyperparameters of each algorithm.

| Algorithm | Values of optimised hyperparameters |
|---|---|
| XGBOOST | max_depth = 20, learning_rate = 0.01, subsample = 0.8999999999999999, colsample_bytree = 0.5, colsample_bylevel = 0.5, n_estimators = 100 |
| Random Forest | n_estimators = 1000, max_depth = 70, max_features = auto, min_samples_leaf = 1, min_samples_split = 2, bootstrap = False |
| Logistic Regression | max_iter = 100, dual = False |
| Support Vector Classification | gamma = 0.001, C = 1 |
| ANN | solver = sgd, learning_rate = adaptive, hidden_layer_sizes= (10,10,10), alpha = 0.0001, activation = relu |
| Decision Tree | min_samples_leaf = 20, max_depth = 2, criterion = gini |
| Gradient Boosting Classification | n_estimators = 500, max_depth = 1, learning_rate = 0.01 |

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

In our work, if two features have a Pearson correlation coefficient greater than 0.8, they are considered to be highly correlated features. The list of correlated features can be found in the Supplementary Information file.

After the removal of correlated features, we are left with a total of 118 features. Then, we used ten-fold cross-validation to test the prediction accuracy of seven machine learning algorithms which include XGBOOST (eXtreme Gradient BOOST) classier, logistic regression, support vector classifier, artificial neural network, decision tree, random forest, and gradient boosting classifier. The XGBOOST algorithm deserves some discussion. It is based on the Gradient Boosting Decision Tree algorithm. Boosting is an ensemble method in which new models are incorporated to rectify the errors which result from existing models. Models are added until the point when further improvement becomes marginal. In the gradient boosting approach, new models calculate the errors of the existing models. These errors are then added, and the final prediction is made. This approach is known as gradient boosting because the gradient descent algorithm is utilised to minimise the loss during the incorporation of new models. The XGBOOST algorithm is faster as compared to other algorithms based on gradient boosting. XGBOOST has many attractive attributes like Block Structure which enables parallelisation of tree construction, Continued Training which allows people to use an already fitted model on new data etc. [42].

Before performing ten-fold cross-validation, the hyperparameters of all these algorithms are optimised using the RandomizedSearchCV technique. Hyperparameter optimisation is essential for maximising accuracy. RandomizedSearchCV is a commonly used hyperparameter optimisation technique that is available in the 'model selection' module of the scikit learn library. In the RandomizedSearchCV technique, a set

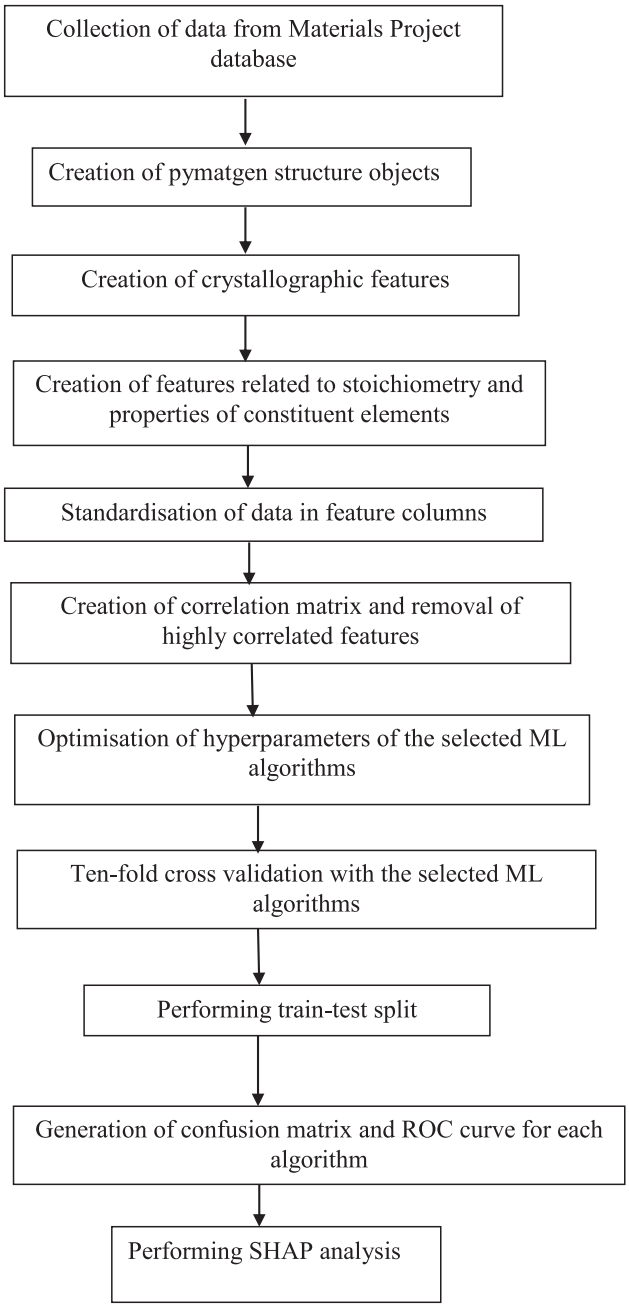


**Fig. 1.** An illustration of the methodology followed in this work.

of values is specified for each important hyperparameter of the algorithm which needs to be optimised. Subsequently, different combinations of hyperparameter values are selected and the performance of the algorithm is measured. The combination of hyperparameter values which gives the best performance is the optimised combination. The total number of combinations which are tried out is determined by the 'n_iter' parameter. Table 1 shows the hyperparameters which were optimised for each algorithm. The values of the optimised hyperparameters have been shown in Table 2. Once hyperparameter optimisation is complete, ten-fold cross validation is performed for the seven algorithms previously mentioned.

After the completion of ten-fold cross-validation, a train-test split is once again performed to generate a confusion matrix for each algorithm. 90% of the rows are included in the training set and the remaining 10% are included in the testing set. As the number of indirect band gap materials in the training set is much higher than the number of direct

**Table 3**
Average classification accuracy of each algorithm after ten-fold cross-validation.

| Algorithm | Average accuracy |
|---|---|
| XGBOOST | 72.8% |
| Random Forest | 73.4% |
| Logistic Regression | 67.3% |
| Support Vector Classification | 73% |
| ANN | 73% |
| Decision Tree | 74.2% |
| Gradient Boosting Classification | 73.1% |

band gap materials, the training set is resampled. Oversampling of the minority class is carried out i.e., duplicate instances of direct band gap materials are created. After the completion of resampling, around 40% of the instances in the training set are direct band gap materials. Each of the seven algorithms mentioned earlier are trained and tested on that particular combination of training set and testing set. Seven different confusion matrices, which show the performance of these algorithms while making predictions on the test set, are generated. For each algorithm, a Receiver Operating Characteristic (ROC) curve is also generated. For generating a ROC curve, the True Positive Rate (TPR) is plotted against the False Positive Rate (FPR) at multiple threshold values. TPR and FPR are defined as follows:

$TPR = TP/(TP + FN)$.

$FPR = FP/(FP + TN)$.

where TP – True Positive, FP – False Positive, TN – True Negative, FN – False Negative.

The efficiency of a classification algorithm can be known by finding the area under a ROC curve. Higher is the value of Area Under Curve (AUC), better is the classifier. A purely random classifier will have AUC equal to 0.5 while a perfect classifier will have AUC equal to 1 [43].

In the end, SHAP analysis is performed to find out which features played the dominant role in the prediction of the nature of band gap. SHAP stands for SHapley Additive exPlanations and it is a technique used to determine the importance of features. Even after the removal of highly correlated features, it is not necessary that all the remaining features will play an important role when an algorithm makes predictions. SHAP analysis can determine the features that actually play an important role in the prediction process. SHAP analysis determines the importance of features on the basis of SHAP values. SHAP values can be understood by the following explanation: let us assume that there are N features in total and S is a subset of the total number of features. Let p(S) represent the contribution of these S features towards the prediction model. If another feature y is added to these S features, the marginal contribution of y is given by p(S U' y) – p(S). If the average of this contribution is taken over all possible ways in which S can be formed, the actual contribution of y can be calculated [44]. Similarly, the contribution of each feature can be calculated. The methodology followed for this work has been presented as a flowchart in Fig. 1.

## 3. Results and discussion

Table 3 shows the average classification accuracy of each algorithm when ten-fold cross-validation is performed. From Table 3, it can be observed that the decision tree classification algorithm gives the highest average accuracy of 74.2% when ten-fold cross-validation is performed. But for our problem, accuracy is not the metric in which we are most interested. We are most interested in the precision of direct band gap prediction i.e., the percentage of materials which are actually direct band gap materials out of the total number of materials which are predicted to have a direct band gap.

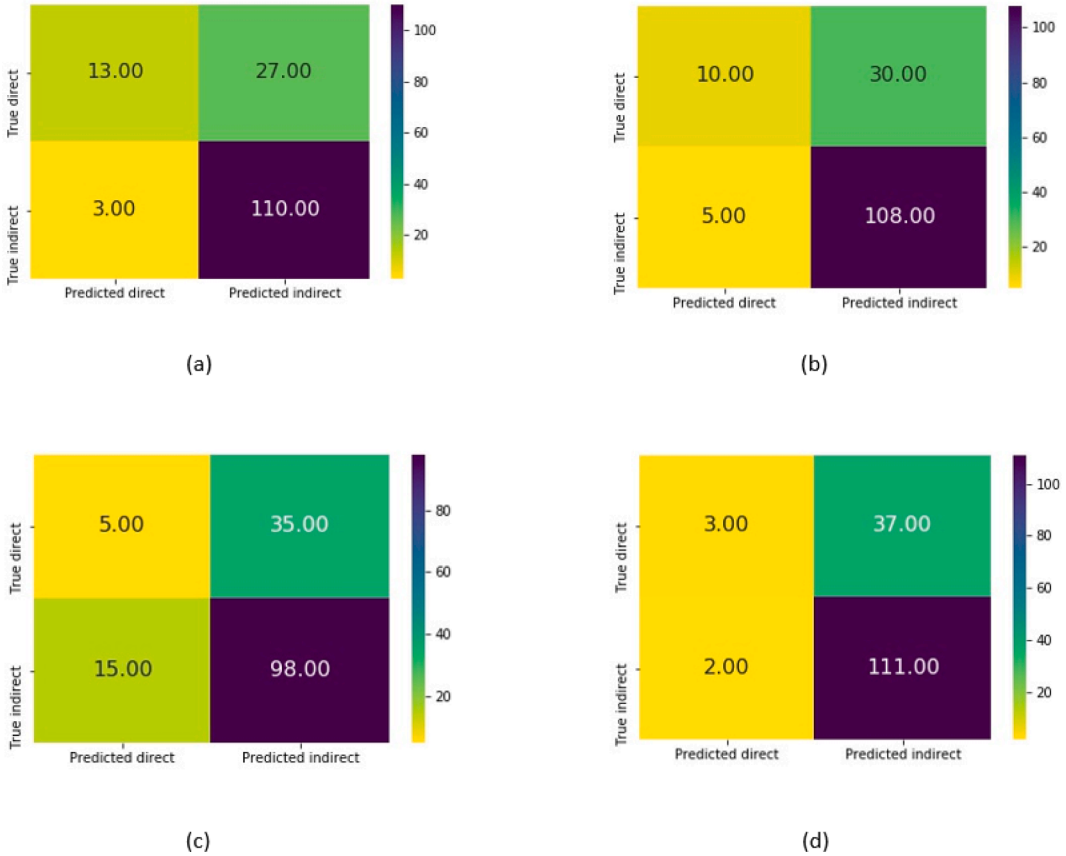


**Fig. 2.** Confusion matrices showing the performance of (a) XGBOOST (b) Random Forest. (c) Logistic Regression (d) Support Vector Classification, a confusion matrix can reveal what percentage of the materials predicted to be direct or indirect band gap materials are actually so.

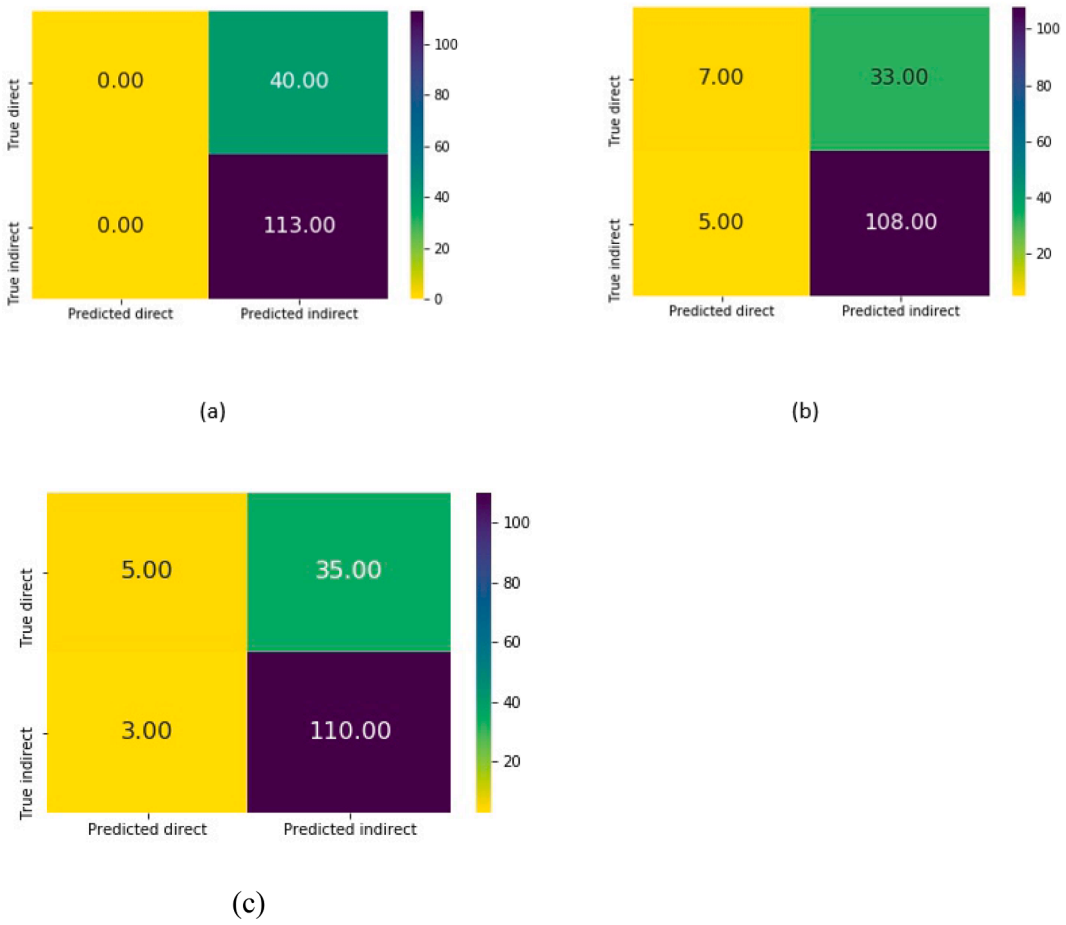


**Fig. 3.** Confusion matrices showing the performance of (a) ANN (b) Decision Tree. (c) Gradient Boosting Classification.

**Table 4**
Precision, recall, and f1-scores for predictions on the test set.

| Algorithm | Precision | | Recall | | F1-score | |
|---|---|---|---|---|---|---|
| | 0 | 1 | 0 | 1 | 0 | 1 |
| XGBOOST | 0.8 | 0.81 | 0.97 | 0.33 | 0.88 | 0.46 |
| Random Forest | 0.78 | 0.67 | 0.96 | 0.25 | 0.86 | 0.36 |
| Logistic Regression | 0.74 | 0.25 | 0.87 | 0.12 | 0.8 | 0.17 |
| Support Vector Classification | 0.75 | 0.6 | 0.98 | 0.07 | 0.85 | 0.13 |
| ANN | 0.74 | 0 | 1 | 0 | 0.85 | 0 |
| Decision Tree | 0.77 | 0.58 | 0.96 | 0.17 | 0.85 | 0.27 |
| Gradient Boosting Classification | 0.76 | 0.62 | 0.97 | 0.12 | 0.85 | 0.21 |

After the completion of ten-fold cross-validation, we performed a train-test split once again to generate a confusion matrix for each algorithm. The confusion matrices for the seven algorithms have been shown in Fig. 2 and Fig. 3.

Confusion matrices are very useful for making a detailed analysis of the performance of ML algorithms in a classification problem. For example, Fig. 3(a) reveals that all the materials in the test set are predicted to be indirect band gap materials by ANN. Thus, ANN is the least suitable algorithm for our problem.

As mentioned earlier, class 0 represents indirect band gap material and class 1 represents direct band gap material. From Table 4, it can be observed that the highest precision for the prediction of class 1 (direct band gap material) is provided by the XGBOOST classifier i.e., 0.81 or 81%. In fact, the XGBOOST classifier is the only algorithm that provides a precision higher than 70%. The XGBOOST classifier provides an f1-score of 0.46 for the prediction of class 1 which is also the highest f1-score provided among all the algorithms. The ROC curves for the seven algorithms have been shown in Figs. 4 and 5.

The area under the ROC curves of the seven algorithms implemented have been given in Table 5. From Table 5, it can be clearly observed that the AUC of XGBOOST classifier is 0.65, the highest among all algorithms. Thus, XGBOOST classifier has the best precision as well as the best AUC. For the XGBOOST classifier, the materials in the test set which are incorrectly classified are analysed. There is nothing directly discernible in common between these materials.

The high precision provided by the XGBOOST algorithm suggests that its use for the prediction of other properties of $ABX_3$ perovskites may yield good results. It will also be interesting to know where XGBOOST stands as compared to the other algorithms which have been used in the past to predict various properties of perovskites. Zhai et al. used the Support Vector Regression (SVR) algorithm to predict the Curie temperature of perovskites with a Root Mean Squared Error (RMSE) of 28.6659 K [45]. Behara et al. used the Light Gradient Boosting Machine (LGBM) classifier to classify the crystal structure of perovskites with an accuracy of 80.3% [44]. They also performed SHAP analysis to determine the features which played the dominant role in the classification of crystal structure. Although Behara et al. performed SHAP analysis in their work, their focus was on the classification of crystal structure of perovskites while our focus is on the prediction of the nature of band gap of perovskites. Pilania et al. used the Kernel Ridge Regression (KRR) algorithm to predict the band gap of double perovskites with an RMSE of 0.371 eV (for primary descriptors) and 0.36 eV (for compound descriptors) [46]. Lu et al. used the Gradient Boosting Regression (GBR) algorithm to predict the band gap of perovskites with a coefficient of determination ($R^2$) of 0.921 [47].

In our work, we have used descriptors containing information about stoichiometry, properties of constituent elements, and crystal structure,

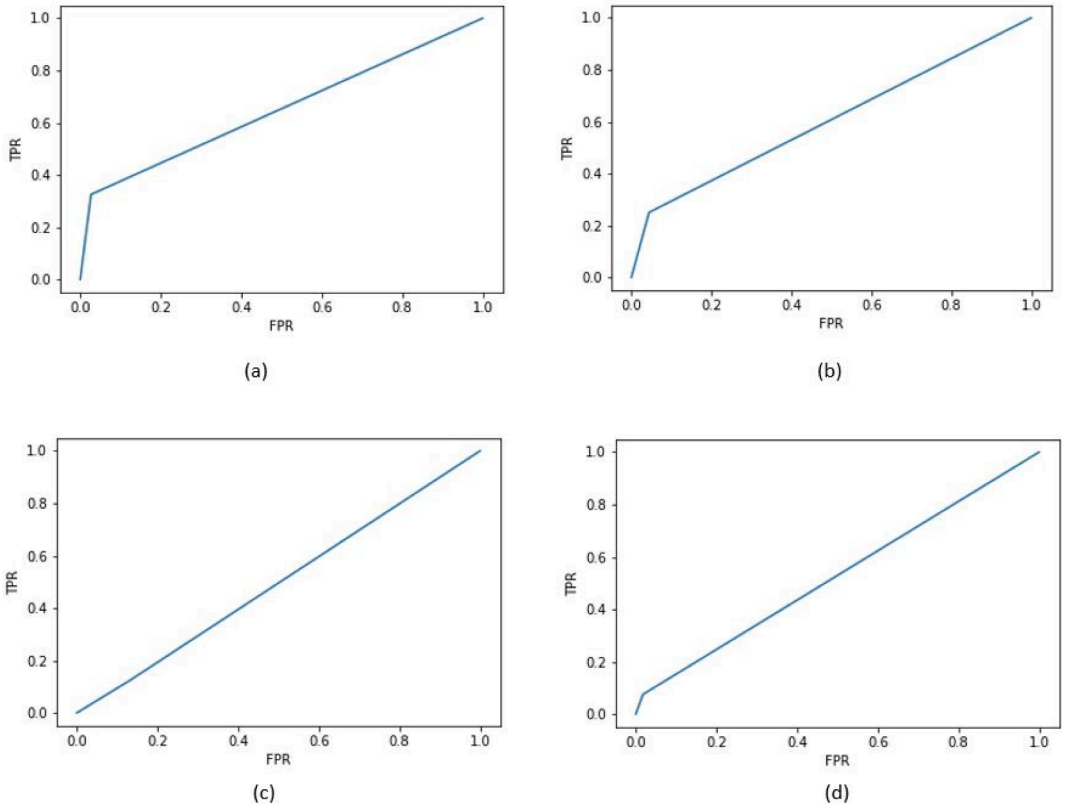


**Fig. 4.** ROC curves of (a) XGBOOST (b) Random Forest (c) Logistic Regression. (d) Support Vector Classification, the area under a ROC curve gives us the efficiency. of a classification algorithm, higher is the area under a ROC curve, higher is the. efficiency of the classification algorithm.

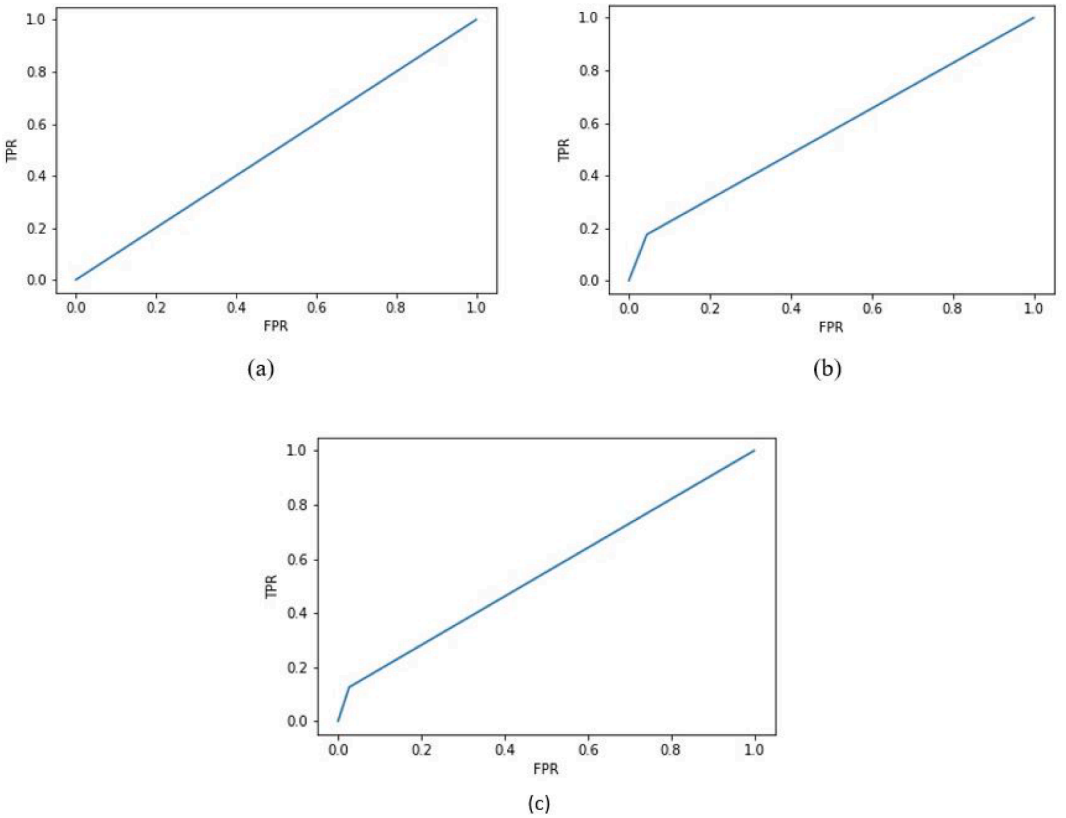


**Fig. 5.** ROC curves of (a) ANN (b) Decision Tree (c) Gradient Boosting Classification.

**Table 5**
Area under ROC curve for each algorithm.

| Algorithm | Area under ROC curve |
|---|---|
| XGBOOST | 0.65 |
| Random Forest | 0.6 |
| Logistic Regression | 0.5 |
| Support Vector Classification | 0.53 |
| ANN | 0.5 |
| Decision Tree | 0.57 |
| Gradient Boosting Classification | 0.55 |

but none of them contain information about electronic charge density. According to Pilania et al.[14], descriptors related to electronic charge density can be used to make accurate predictions of material properties. Hence evaluating the performance of XGBOOST algorithm with electronic charge density descriptors can be an interesting research direction.

As the XGBOOST classifier provides the highest precision for the prediction of direct band gap materials, SHAP analysis is performed only for the XGBOOST classifier. After the SHAP analysis is performed, feature importance graph and SHAP summary plot for the XGBOOST classifier are generated. Although we used 118 features to train the XGBOOST classifier, not all of them played an important role in the prediction of the nature of band gap.

Fig-6 shows the feature importance graph. In this figure, the 20 most important features which played the dominant role in the prediction of the nature of band gap have been arranged from top to bottom in the descending order of their importance. Thus, the most important feature can be found at the top and the least important feature can be found at the bottom.

Along the Y-axis, we find the names of the features and along the X-axis, we find the corresponding mean of the magnitude of SHAP values. The mean of the magnitude of SHAP values of a feature represents the average impact of that feature on the model output. Higher is the mean of the magnitude of SHAP values of a feature, higher is its impact on the final predicted value of the nature of band gap (0 or 1). It can be clearly observed that the 10 most important features are - average deviation neighbor distance variation, minimum neighbor distance variation, average d valence electrons, mean atomic weight, maximum relative bond length, range electronegativity, range atomic radius, mean absolute deviation in relative bond length, mean column and maximum neighbor distance variation.

Fig. 7 shows the SHAP summary plot for the XGBOOST classifier. Along the Y-axis are the names of the features which have been arranged from top to bottom in the descending order of their importance. Along the X-axis are the corresponding SHAP values which represent the impact of a feature on the model output. In the summary plot, each point represents a certain instance (row) of the corresponding feature. The colour of a point represents the value of the corresponding feature, high value being represented by red and low value being represented by blue. A positive SHAP value signifies a higher value of the target property (in our case, class 1) while a negative SHAP value signifies a lower value of the target property (in our case, class 0). The magnitude of the SHAP value is linked to probability. Higher is the magnitude of the SHAP value, higher is the probability of the corresponding class. By observing the SHAP summary plot, the following inferences can be made:

- Lower is the average number of valence d electrons in a compound, higher is the probability of the compound having a direct band gap. The average number of valence d electrons in a compound can be

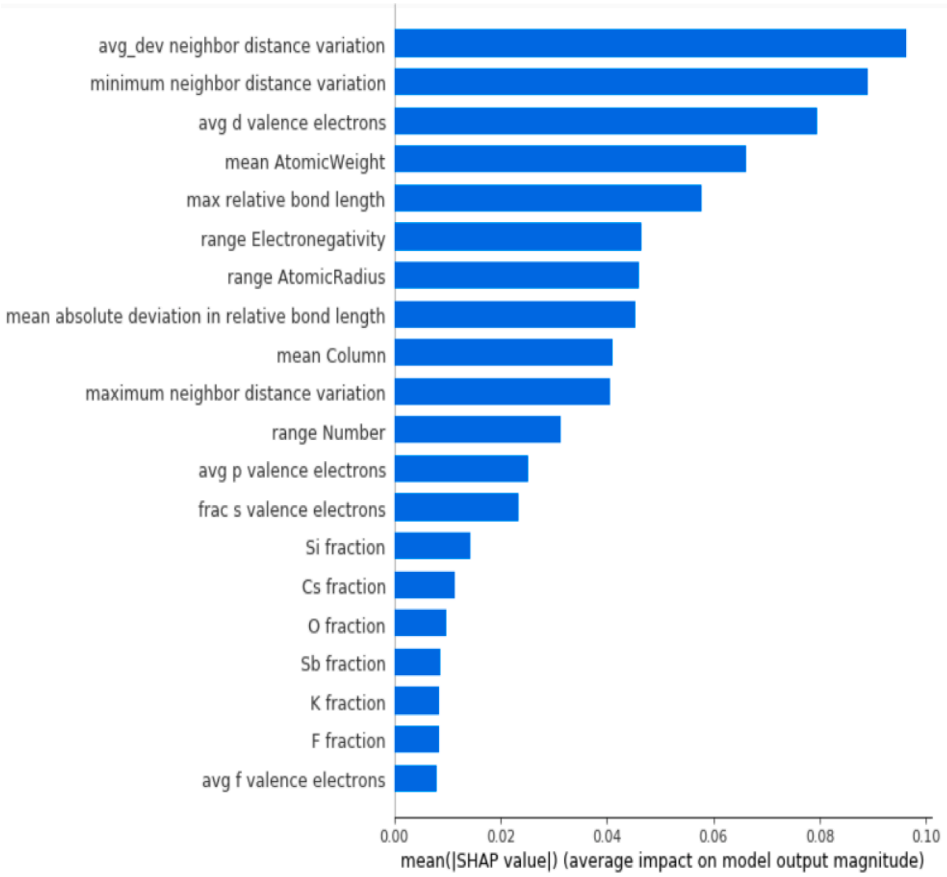


**Fig. 6.** Feature importance graph for the XGBOOST classifier, the importance of features decreases as we go down i.e., the features at the top play the most important role in predicting the nature of band gap.

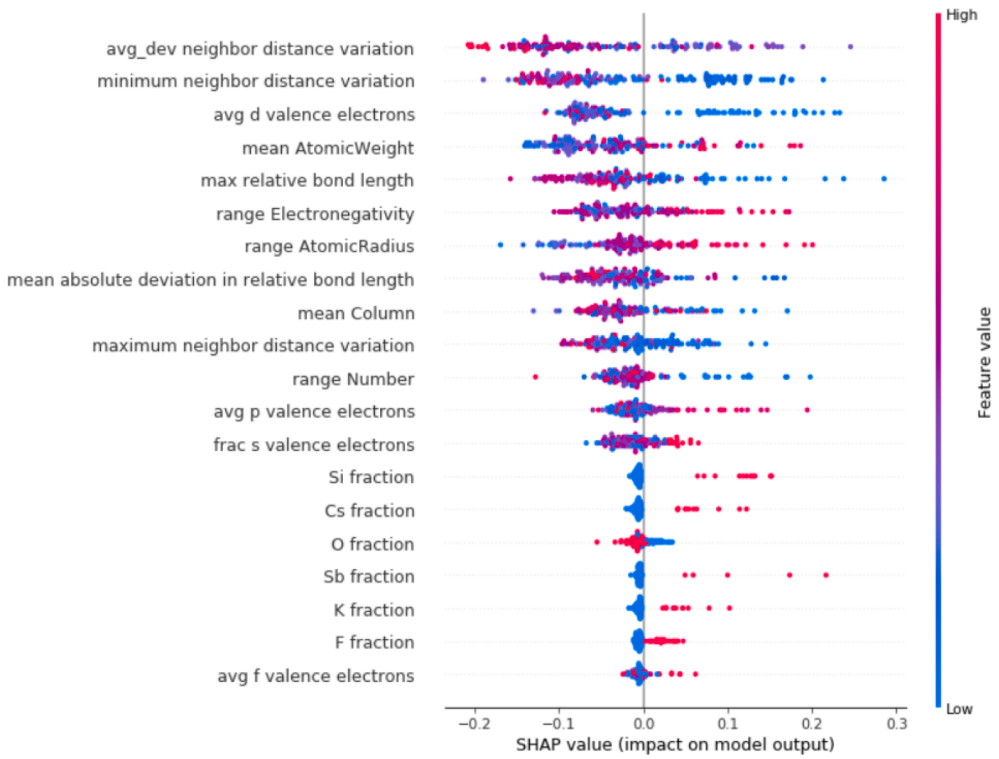


**Fig. 7.** SHAP summary plot for XGBOOST classifier, this plot shows how the values of a certain feature influence the nature of band gap.

obtained by dividing the total number of valence d electrons in one formula unit of the compound by the total number of atoms in one formula unit. The d electrons present in the penultimate shell of an atom are considered as valence electrons. The first 20 elements of the periodic table (H to Ca) do not contain d electrons. So, for these elements, the number of valence d electrons is 0. Alkali metals and alkaline earth metals don't contain d electrons in the penultimate shell (although they may have d electrons in deeper shells). So, the number of valence d electrons is 0 for these elements as well. However, transition metals and elements belonging to groups IIIA to VIIIA which have atomic number greater than 20 contain d electrons in their penultimate shell. For these categories of elements, the number of valence d electrons is not 0. Thus, if a compound does not contain any transition metal and any element belonging to groups IIIA to VIIIA which has atomic number greater than 20, there is a relatively higher probability that it will have a direct band gap. Transition metal compounds also suffer from the problem of small polaron formation [48] which typically decreases the mobility of electrons and holes. This happens because transition metal ions exhibit multiple oxidation states which leads to the localisation of electrons and holes. Hence, the presence of transition metals introduces multiple problems.

- Higher is the range of electronegativity i.e., the difference between maximum electronegativity and minimum electronegativity for the constituent elements of a compound, higher is the probability of the compound having a direct band gap.
- The presence of Si, Cs, Sb, K, or F in a compound increases the probability of the compound having a direct band gap.
- The presence of oxygen increases the probability of the compound having an indirect band gap.

The inferences mentioned above are important from a design standpoint. These inferences can help a researcher with the selection of constituent elements for the $ABX_3$ perovskite so as to maximise the chances of getting a direct band gap material. Gladkikh et al. [49] used machine learning to predict the magnitude of band gap of $ABX_3$ perovskites and then identify the factors which determine the magnitude of band gap. According to their model, the magnitude of band gap is mostly determined by electronegativities, electron affinities, ionization energies and atomic radii of the constituent elements. Figs. 6 and 7 reveal that the range of electronegativity and the range of atomic radii are among the 10 most important features which determine the nature of band gap. Thus, electronegativities and atomic radii of constituent elements play an important role in determining both magnitude and nature of band gap. Our work also reveals factors other than electronegativity and atomic radii which determine the nature of band gap. Thus, our results are consonant with and add to the insights offered by Gladkikh et al.

## 4. Conclusion

In this work, we demonstrate that the XGBOOST classifier can predict direct band gap $ABX_3$ perovskites with a precision of 81%, which is good enough for enabling practical exploitation of the algorithm. When ten-fold cross-validation is used on the dataset, the XGBOOST classifier classifies the nature of band gap with an average accuracy of 72.8%. As we use a diverse dataset containing oxides, sulphides, selenides, tellurides, chlorides, iodides etc.; our model can be used for the prediction of the nature of band gap of non-oxide perovskites.

SHAP analysis performed in the end yields the factors which influence the nature of band gap of perovskites. The absence of transition metals and elements belonging to groups IIIA to VIIIA which have atomic number greater than 20, higher electronegativity difference, absence of oxygen, and the presence of Si, Cs, Sb, K, or F in the perovskite increase the probability of the perovskite having a direct band gap. The presence of a transition metal in a perovskite decreases the probability of possession of a direct band gap and leads to low electron and hole mobilities at the same time. Thus, transition metals should be ideally avoided while synthesising perovskites for light-harvesting applications. The insights offered here are thus design insights that can aid targeted discovery of new $ABX_3$ perovskites which possess direct band gaps. This targeted approach can reduce our dependence on random hit

and trial experiments, and also minimise the amount of ab initio computations needed. This is hence consonant with the ultimate aim of materials informatics, which is accelerated materials selection and design.

## CRediT authorship contribution statement

**Smarak Rath:** Conceptualization, Data curation, Methodology, Visualization, Investigation, Formal analysis, Writing – original draft, Writing – review & editing. **G. Sudha Priyanga:** Conceptualization, Investigation, Formal analysis, Writing – review & editing. **N. Nagappan:** Methodology, Visualization, Writing – review & editing. **Tiju Thomas:** Conceptualization, Supervision, Investigation, Formal analysis, Writing – review & editing.

## 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.

*Acknowledgements*

Smarak Rath gratefully acknowledges the Half Time Research Assistantship (HTRA) provided by the Ministry of Human Resource Development (MHRD), Government of India. We also thank the Department of Metallurgical and Materials Engineering, Indian Institute of Technology Madras. The research did not receive any specific grant from funding agencies in the public, commercial or not-for-profit sectors.

*Data availability*

The raw and processed data required to reproduce these findings and the code that we have used for our work are available to download from https://github.com/smarakrath/MI-2021. The file labelled as 'Supplementary Information' contains a description of all other files uploaded in our GitHub repository. It also contains the complete list of descriptors used for our work.

## References

[1] S. Luo, T. Li, X. Wang, M. Faizan, L. Zhang, High-throughput computational materials screening and discovery of optoelectronic semiconductors, Wiley Interdiscip. Rev. Comput. Mol. Sci. 11 (2021), e1489, https://doi.org/10.1002/wcms.1489.
[2] J. Pan, Q. Yan, Data-driven material discovery for photocatalysis: a short review, J. Semicond. 39 (7) (2018) 071001.
[3] W.J. Yin, J.H. Yang, J. Kang, Y. Yan, S.H. Wei, Halide perovskite materials for solar cells: A theoretical review, J. Mater. Chem. A. 3 (2015) 8926–8942, https://doi.org/10.1039/c4ta05033a.
[4] R.G. Stair, R. Johnston, T.C. Bagg, Spectral distribution of energy from the sun, J. Res. Natl. Bur. Stand. 53 (1954) 113–119, https://doi.org/10.6028/jres.053.014.
[5] S.A. Olaleru, J.K. Kirui, D. Wamwangi, K.T. Roro, B. Mwakikunga, Perovskite solar cells: The new epoch in photovoltaics, Sol. Energy. 196 (2020) 295–309, https://doi.org/10.1016/j.solener.2019.12.025.
[6] F. Liu, Q. Zeng, J. Li, X. Hao, A. Ho-Baillie, J. Tang, M.A. Green, Emerging inorganic compound thin film photovoltaic materials: Progress, challenges and strategies, Mater. Today. 41 (2020) 120–142, https://doi.org/10.1016/j.mattod.2020.09.002.
[7] A. Rocket, The Materials Science of Semiconductors, Springer Science + Business Media, LLC, New York, 2008.
[8] J.G. Lee, Computational Materials Science (An Introduction), 2nd ed., CRC Press, Taylor & Francis Group, LLC, Boca Raton, 2017.
[9] A. Agrawal, A. Choudhary, Perspective: Materials informatics and big data: Realization of the "fourth paradigm" of science in materials science, APL Mater. 4 (5) (2016) 053208.
[10] Y. Liu, T. Zhao, W. Ju, S. Shi, S. Shi, S. Shi, Materials discovery and design using machine learning, J. Mater. 3 (2017) 159–177, https://doi.org/10.1016/j.jmat.2017.08.002.
[11] K. Fujimura, A. Seko, Y. Koyama, A. Kuwabara, I. Kishida, K. Shitara, C.A.J. Fisher, H. Moriwake, I. Tanaka, Accelerated materials design of lithium superionic conductors based on first-principles calculations and machine learning algorithms, Adv. Energy Mater. 3 (2013) 980–985, https://doi.org/10.1002/aenm.201300060.
[12] A. Alzghoul, A. Alhalaweh, D. Mahlin, C.A.S. Bergström, Experimental and computational prediction of glass transition temperature of drugs, J. Chem. Inf. Model. 54 (2014) 3396–3403, https://doi.org/10.1021/ci5004834.
[13] F. Gharagheizi, P. Ilani-Kashkouli, A.H. Mohammadi, A group contribution method for estimation of glass transition temperature ionic liquids, Chem. Eng. Sci. 81 (2012) 91–105, https://doi.org/10.1016/j.ces.2012.06.052.
[14] G. Pilania, C. Wang, X. Jiang, S. Rajasekaran, R. Ramprasad, Accelerating materials property predictions using machine learning, Sci. Rep. 3 (2013) 2810, https://doi.org/10.1038/srep02810.
[15] N. Castin, J.R. Fernández, R.C. Pasianot, Predicting vacancy migration energies in lattice-free environments using artificial neural networks, Comput. Mater. Sci. 84 (2014) 217–225, https://doi.org/10.1016/j.commatsci.2013.12.016.
[16] Z.W. Zhao, M. Del Cueto, Y. Geng, A. Troisi, Effect of Increasing the Descriptor Set on Machine Learning Prediction of Small Molecule-Based Organic Solar Cells, Chem. Mater. 32 (2020) 7777–7787, https://doi.org/10.1021/acs.chemmater.0c02325.
[17] K. Choudhary, M. Bercx, J. Jiang, R. Pachter, D. Lamoen, F. Tavazza, Accelerated Discovery of Efficient Solar Cell Materials Using Quantum and Machine-Learning Methods, Chem. Mater. 31 (2019) 5900–5908, https://doi.org/10.1021/acs.chemmater.9b02166.
[18] L. Wei, X. Xu, Gurudayal, J. Bullock, J.W. Ager, Machine Learning Optimization of p-Type Transparent Conducting Films, Chem. Mater. 31 (18) (2019) 7340–7350.
[19] Y. Zhang, X. Xu, Machine learning the magnetocaloric effect in manganites from lattice parameters, Appl. Phys. A Mater. Sci. Process. 126 (2020) 341, https://doi.org/10.1007/s00339-020-03503-8.
[20] C. Kim, G. Pilania, R. Ramprasad, Machine Learning Assisted Predictions of Intrinsic Dielectric Breakdown Strength of ABX3 Perovskites, J. Phys. Chem. C. 120 (2016) 14575–14580, https://doi.org/10.1021/acs.jpcc.6b05068.
[21] L. Xu, L. Wencong, P. Chunrong, S. Qiang, G. Jin, Two semi-empirical approaches for the prediction of oxide ionic conductivities in ABO3 perovskites, Comput. Mater. Sci. 46 (2009) 860–868, https://doi.org/10.1016/j.commatsci.2009.04.047.
[22] K. Takahashi, L. Takahashi, I. Miyazato, Y. Tanaka, Searching for Hidden Perovskite Materials for Photovoltaic Systems by Combining Data Science and First Principle Calculations, ACS Photonics 5 (2018) 771–775, https://doi.org/10.1021/acsphotonics.7b01479.
[23] S. Behara, G.S. Priyanga, T. Thomas, Strain-induced effects in the electronic and optical properties of Na0.5Bi0.5TiO3: An ab-initio study, Mater. Today Commun. 24 (2020), 101348, https://doi.org/10.1016/j.mtcomm.2020.101348.
[24] G. Sudha Priyanga, T. Thomas, Effective mass and optical properties of orthorhombic Al1−xInxFeO3 perovskite: An ab-initio study, Comput. Mater. Sci. 159 (2019) 222–227, https://doi.org/10.1016/j.commatsci.2018.12.012.
[25] G. Sudha Priyanga, T. Thomas, Direct band gap narrowing and light-harvesting-potential in orthorhombic In-doped-AlFeO3 perovskite: A first principles study, J. Alloys Compd. 750 (2018) 312–319, https://doi.org/10.1016/j.jallcom.2018.03.388.
[26] J. Li, B. Pradhan, S. Gaur, J. Thomas, Predictions and Strategies Learned from Machine Learning to Develop High-Performing Perovskite Solar Cells, Adv. Energy Mater. 9 (2019) 1901891, https://doi.org/10.1002/aenm.201901891.
[27] L. Ward, A. Dunn, A. Faghaninia, N.E.R. Zimmermann, S. Bajaj, Q. Wang, J. Montoya, J. Chen, K. Bystrom, M. Dylla, K. Chard, M. Asta, K.A. Persson, G. J. Snyder, I. Foster, A. Jain, Matminer: An open source toolkit for materials data mining, Comput. Mater. Sci. 152 (2018) 60–69, https://doi.org/10.1016/j.commatsci.2018.05.018.
[28] S.P. Ong, W.D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K.A. Persson, G. Ceder, Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis, Comput. Mater. Sci. 68 (2013) 314–319, https://doi.org/10.1016/j.commatsci.2012.10.028.
[29] F. Pedregosa, et al., Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.
[30] T. Chen, C. Guestrin, in: XGBoost: A Scalable Tree Boosting System, ACM, New York, NY, USA, 2016, pp. 785–794, https://doi.org/10.1145/2939672.2939785.
[31] S.M. Lundberg, G. Erion, H. Chen, A. DeGrave, J.M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, S.-I. Lee, From local explanations to global understanding with explainable AI for trees, Nat. Mach. Intell. 2 (2020) 56–67, https://doi.org/10.1038/s42256-019-0138-9.
[32] Data structures for statistical computing in python, McKinney, Proceedings of the 9th Python in Science Conference 445 (2010).
[33] C.R. Harris, K.J. Millman, S.J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N.J. Smith, R. Kern, M. Picus, S. Hoyer, M.H. van Kerkwijk, M. Brett, A. Haldane, J.F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, T.E. Oliphant, Array programming with NumPy, Nature 585 (2020) 357–362, https://doi.org/10.1038/s41586-020-2649-2.
[34] Materials Project. https://materialsproject.org/ (accessed 8 April 2022).
[35] Citrination. https://citrination.com/ (accessed 8 April 2022).
[36] Materials Data Facility. https://materialsdatafacility.org/ (accessed 8 April 2022).
[37] Materials Platform for Data Science. https://mpds.io/ (accessed 8 April 2022).
[38] L. Ward, R. Liu, A. Krishna, V.I. Hegde, A. Agrawal, A. Choudhary, C. Wolverton, Including crystal structure attributes in machine learning models of formation energies via Voronoi tessellations, Phys. Rev. B. 96 (2017), 024104, https://doi.org/10.1103/PhysRevB.96.024104.
[39] B. Meredig, A. Agrawal, S. Kirklin, J.E. Saal, J.W. Doak, A. Thompson, K. Zhang, A. Choudhary, C. Wolverton, Combinatorial screening for new materials in

unconstrained composition space with machine learning, Phys. Rev. B 89 (2014), 094104, https://doi.org/10.1103/PhysRevB.89.094104.
[40] K. Rajan, Materials informatics, Mater. Today. 8 (2005) 38–45, https://doi.org/10.1016/S1369-7021(05)71123-8.
[41] Real Python; Numpy, SciPy, and Pandas: Correlation With Python. https://realpython.com/numpy-scipy-pandas-correlation-python/ (accessed 9 April 2022).
[42] Machine Learning Mastery; A Gentle Introduction to XGBoost for Applied Machine Learning. https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/ (accessed 9 April 2022).
[43] A. Géron, Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow, 2nd ed., O' Reilly Media Inc, Sebastopol, 2019.
[44] S. Behara, T. Poonawala, T. 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.
[45] X. Zhai, M. Chen, W. Lu, Accelerated search for perovskite materials with higher Curie temperature based on the machine learning methods, Comput. Mater. Sci. 151 (2018) 41–48, https://doi.org/10.1016/j.commatsci.2018.04.031.
[46] G. Pilania, A. Mannodi-Kanakkithodi, B.P. Uberuaga, R. Ramprasad, J. E. Gubernatis, T. Lookman, Machine learning bandgaps of double perovskites, Sci. Rep. 6 (2016) 19375, https://doi.org/10.1038/srep19375.
[47] S. Lu, Q. Zhou, L. Ma, Y. Guo, J. Wang, Rapid Discovery of Ferroelectric Photovoltaic Perovskites and Material Descriptors via Machine Learning, Small Methods. 3 (2019) 1900360, https://doi.org/10.1002/smtd.201900360.
[48] S. Lany, Semiconducting transition metal oxides, J. Phys. Condens. Matter. 27 (28) (2015) 283203.
[49] V. Gladkikh, D.Y. Kim, A. Hajibabaei, A. Jana, C.W. Myung, K.S. Kim, Machine Learning for Predicting the Band Gaps of ABX3 Perovskites from Elemental Properties, J. Phys. Chem. C. 124 (2020) 8905–8918, https://doi.org/10.1021/acs.jpcc.9b11768.