Impact of SMOTE Oversampling on Classifying Band Gap Types in Imbalanced ABO_3 Perovskite Oxides

Desvita Maharani[1], **Muhamad Akrom**[2*], **Johana Oktavia Ramadhani**[1], **& Aliyah Zahratu Rizqi**[1]
[1]Study Program in Informatic, Dian Nuswantoro University, Indonesia
[2]Research Center for Quantum Computing and Materials Informatics, Dian Nuswantoro University, Indonesia
*Corresponding Author: m.akrom@dsn.dinus.ac.id

Received: 3rd March 2026; **Accepted**: 24th April 2026; **Published**: 30th April 2026
DOI: https://dx.doi.org/10.29303/jpft.v12i1.11479

Abstract - *This study investigates the impact of the Synthetic Minority Over-sampling Technique (SMOTE) on the classification of direct and indirect band gap types in imbalanced ABO_3 perovskite oxide datasets. In the dataset used, the direct band gap class constitutes approximately 84% of the samples, while the indirect class represents only 16%, leading conventional classification models to become biased toward the majority class. To address this issue, SMOTE was employed to balance the class distribution, and its performance was evaluated using several machine learning algorithms, including Multi-Layer Perceptron (MLP), Extra Trees, CatBoost, and Gradient Boosting. Model performance was assessed using 5-fold stratified cross-validation, with particular emphasis on F1-macro and recall metrics to ensure adequate evaluation of the minority class. The results show that although SMOTE did not significantly improve overall accuracy (baseline: 0.89; SMOTE: 0.88), it enhanced the models' ability to recognize the minority class. Notable improvements in F1-macro were observed, increasing from 0.76 to 0.78 for MLP and from 0.75 to 0.78 for CatBoost. These findings highlight the importance of using F1-macro as a more informative evaluation metric than accuracy for imbalanced datasets and provide methodological insights for developing more robust predictive models in materials informatics.*

Keywords*: SMOTE; class imbalance; perovskite oxide ABO_3; band gap classification (Direct vs Indirect)*

INTRODUCTION

Perovskite oxides with the general formula ABO_3 are an important class of materials that play a central role in the development of modern energy and electronic technologies. However, not all ABO_3 compounds naturally form perovskite structures, making reliable prediction methods essential for identifying stable perovskite phases (Rahman et al., 2025).

Perovskite materials exhibit a wide range of multifunctional properties, including piezoelectricity, ferroelectricity, and superconductivity (Dawa & Sajjadi, 2024). Their compositional flexibility and tunable electronic properties, achieved through elemental substitution at the A and B sites, have enabled extensive applications in solar cells, electrochemical catalysts, sensors, and semiconductor devices (D. Kim et al., 2022; Liu et al., 2021). Among the key properties governing the performance of perovskite oxides is the band gap, defined as the energy difference between the valence and conduction bands. This property strongly influences light absorption, charge transport, and the optoelectronic behavior of materials (Hoye et al., 2022; Shi et al., 2021). Consequently, accurate classification and prediction of band gap characteristics are essential for the rational design of functional materials.

Conventional electronic structure prediction methods based on density functional theory (DFT) face several limitations, including high computational cost and restricted accuracy arising from approximations in exchange–correlation

functionals (Akrom et al., 2024; Zhao et al., 2024). As an alternative, machine learning (ML) approaches have gained considerable attention due to their computational efficiency and ability to capture complex relationships between physicochemical descriptors and material properties, such as band gaps, at significantly higher speeds than first-principles methods (Wang et al., 2021; Zhang et al., 2023).

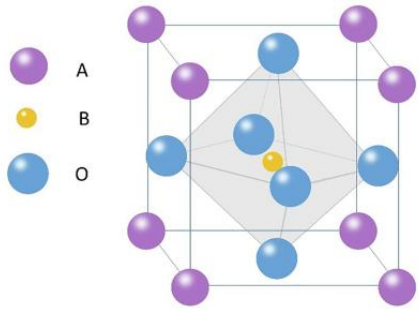

Figure 1. The structure of perovskite ABO_3 is cubic, where the purple spheres represent cation A, the yellow spheres represent cation B, and the blue spheres represent oxygen anions that form octahedrons (Rahman et al., 2025).

By leveraging the continuously expanding materials database, machine learning (ML) has significant potential to accelerate the exploration and screening of candidate materials prior to experimental validation (Wang et al., 2021). However, the application of ML in materials informatics remains challenged by issues related to data quality and dataset characteristics. Real-world materials datasets frequently exhibit imbalanced class distributions due to experimental limitations, overrepresentation of specific material types, and biases in data collection and publication processes.

In the classification of perovskite oxide band gap types, this issue is reflected in the unequal distribution between direct and indirect band gap classes. The indirect class constitutes only approximately 16% of the total dataset, whereas the direct class accounts for nearly 84%. Such imbalance may cause machine learning models to develop decision boundaries that are heavily biased toward the majority class. Consequently, minority-class samples are often underrepresented during model training, reducing the model's ability to generalize effectively. Dataset balancing techniques can mitigate this issue by providing a more proportional representation of each class during training, thereby reducing majority-class bias (Mujahid et al., 2024).

Most previous studies have relied primarily on global accuracy as the main performance metric, despite its limited reliability for imbalanced classification problems. In highly imbalanced datasets, a model may achieve high accuracy simply by consistently predicting the majority class (Owusu-Adjei et al., 2023), while failing to correctly identify minority-class samples that may possess important functional significance, such as indirect band gap materials in optoelectronic applications.

To address this challenge, imbalance-handling strategies are required to improve model sensitivity toward minority classes without compromising generalization performance. Imbalanced learning methods are specifically designed to overcome classification problems caused by insufficient minority-class samples (Xu et al., 2023). One of the most widely used approaches is the Synthetic Minority Over-sampling Technique (SMOTE), which generates synthetic minority samples through interpolation between neighboring data points (Mukherjee & Khushi, 2021). Although SMOTE has demonstrated effectiveness in improving minority-class performance, it also presents several limitations, including the generation of synthetic samples that may not accurately reflect the original data distribution and the possibility of suboptimal sample synthesis (Ramadhan, 2025). Moreover, uncontrolled preprocessing procedures, such as applying standardization and oversampling before data partitioning, may introduce data

leakage and lead to overly optimistic performance estimates (Park et al., 2024).

Several studies have explored strategies for handling class imbalance in machine learning-based material classification. Sudha et al., (2022) applied a simple random oversampling approach for ABO_3 perovskite band gap classification using Random Forest; however, their evaluation relied solely on accuracy without explicitly addressing imbalance-related methodological concerns. In addition, undersampling-based techniques such as Tomek Links and Edited Nearest Neighbours (ENN) have been investigated for imbalanced classification tasks. Nevertheless, these methods primarily remove ambiguous majority-class samples near decision boundaries without enhancing the representation of the minority class itself. In contrast, SMOTE enriches minority-class representation by generating synthetic samples through interpolation within the minority-class feature space, making it particularly suitable for datasets in which minority samples are insufficient to adequately define class boundaries. To the best of the authors' knowledge, no previous study has systematically evaluated the impact of SMOTE on ABO_3 perovskite band gap classification using F1-macro as the primary evaluation metric across multiple machine learning algorithms within a controlled leakage-free pipeline.

Based on these considerations, this study aims to comprehensively analyze the impact of SMOTE on the classification of band gap types in ABO_3 perovskite compounds. Particular emphasis is placed on balancing overall predictive accuracy and minority-class recognition performance. Furthermore, this study compares the performance of several machine learning algorithms, namely Multilayer Perceptron (MLP), Extra Trees, CatBoost, and Gradient Boosting, in handling synthetic data generated through oversampling. Model evaluation is conducted using stratified cross-validation to preserve class proportions within each fold. To prevent data leakage, standardization and SMOTE are applied exclusively to the training data in each validation fold. Through this framework, the study aims to provide methodological insights for developing more reliable machine learning-based predictive models to support rational and efficient materials discovery and design.

RESEARCH METHOD

This study employed a quantitative research design based on computational experiments to compare the performance of classification models under two experimental settings: a baseline configuration without imbalance handling and a configuration incorporating the Synthetic Minority Over-sampling Technique (SMOTE) on the training data. SMOTE is an interpolation-based oversampling method that generates synthetic minority-class samples within the feature space by interpolating between existing minority samples and their nearest neighbors. Unlike direct duplication techniques, SMOTE expands the minority-class distribution more continuously, thereby improving its statistical representation during model training (Chawla et al., 2002; Elreedy et al., 2024). The experimental workflow was structured as a systematic pipeline consisting of data understanding, preprocessing, model training, and performance evaluation stages. A flowchart illustrating the overall experimental procedure is presented in Figure 2.

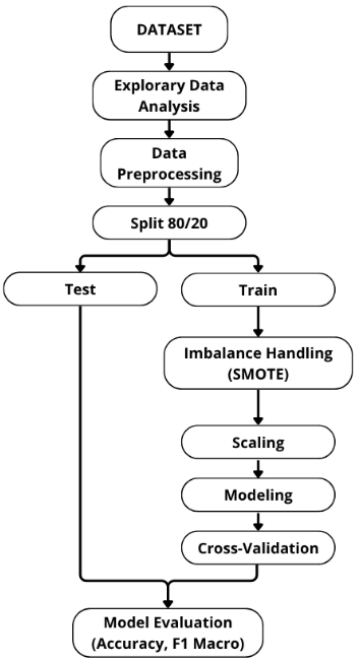

Figure 2. Flowchart

The dataset used in this study consists of secondary tabular data representing perovskite oxide materials with the general formula ABO_3. The initial dataset contained 5,940 entries; after duplicate removal, a total of 3,469 unique samples were retained for all experiments. Each sample was characterized using six physicochemical descriptors, including the ionic radii of elements A and B (RA and RB), the electronegativities of elements A and B (ENA and ENB), the average electronegativity ratio (ENR), and the average ionic character of the material. These features were selected due to their strong relevance to the crystal structure and electronic properties of perovskite materials.

Table 1. Dataset

RA	1,05	1,12	0,8	0,83
RB	1,18	0,27	0,71	0,78
ENA	1,25	1,1	1,66	1,3
ENB	0,95	2,04	1,3	1,83
ENR	-3,306393	-1,117357	-1,935071	-2,0316787
Avg ionic char	0,1792422681	0,1438865627	0,1487374142	0,1417526085
Band gap	0	0	0	0

The target variable corresponds to the band gap type, categorized into two classes: direct band gap and indirect band gap. The dataset exhibits a significant class imbalance, with the direct band gap class accounting for approximately 84% of the samples, while the indirect band gap class represents only 16%. Class imbalance refers to a disproportionate distribution of samples across target classes, in which one class substantially outnumbers the others (Buda et al., 2018). In this study, the dominance of the direct band gap class may bias the learning process toward majority-class patterns.

Such imbalance can lead machine learning models to construct decision boundaries that are overly influenced by the majority class, since the optimization process is primarily driven by classes with larger sample sizes. Consequently, the model may exhibit reduced sensitivity in recognizing minority-class characteristics, despite achieving high overall accuracy (Joloudari et al., 2023; B. Kim & Kim, 2020). Prior to modeling, data exploration and preprocessing were performed to examine dataset characteristics and ensure data quality. Preliminary analysis indicated the absence of dominant linear relationships between the input features and the target variable, suggesting that the class separation pattern is inherently non-linear.

Table 2. Summary of Dataset and Class Distribution

Description	Value
Initial Entry Amount	5.940
Final Sample Size	3.469
Number of Features	6
Majority Class (Direct)	2.775 (84%)
Minority Class (Indirect)	694 (16%)

Data preprocessing involved the removal of duplicate entries to reduce potential bias, while extreme values were retained because they were considered to

represent physically meaningful material variations. All numerical features were subsequently standardized using StandardScaler to normalize feature scales, particularly for models sensitive to differences in feature magnitude (Assegie et al., 2023).

Model evaluation was performed using a 5-fold stratified cross-validation scheme to preserve class proportions in each fold and obtain stable, representative performance estimates (Szeghalmy & Fazekas, 2023). To address class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was applied exclusively to the training data within each cross-validation iteration. This approach ensured balanced class distributions during training while preventing modifications to the test data and minimizing the risk of data leakage (Demircioğlu, 2024). The overall experimental workflow, including fold partitioning, feature standardization, SMOTE application, and model training, is illustrated in Figure 2.

Four machine learning algorithms were employed in this study: Multilayer Perceptron (MLP), Extra Trees, CatBoost, and Gradient Boosting. These models were intentionally selected to represent distinct learning architectures. MLP represents a neural network-based approach capable of capturing complex non-linear relationships through layered feature transformations. In contrast, Extra Trees, CatBoost, and Gradient Boosting represent ensemble-based approaches with different learning mechanisms, including bagging-based parallel tree construction in Extra Trees and sequential boosting strategies in CatBoost and Gradient Boosting. This architectural diversity enables a more comprehensive evaluation of how different learning paradigms respond to synthetic data generated by SMOTE under imbalanced class conditions.

The MLP model consisted of two hidden layers containing 64 and 32 neurons, respectively, and utilized the ReLU activation function with the Adam optimizer. Training was performed for a maximum of 500 iterations. The ensemble-based models were implemented using their default parameter settings to maintain consistency and fairness in performance comparison. All experiments were conducted using a fixed random seed to ensure reproducibility. The computational experiments were implemented in Python 3.x using several libraries, including scikit-learn, catboost, imbalanced-learn, pandas, and numpy, within a cloud-based computing environment.

Table 3. Model and Configuration Summary

Model	Main Configuration
MLP	2 hidden layer (64, 32), ReLU, Adam
ExtraTrees	Default
CatBoost	Default
Gradient Boosting	Default

Model performance is evaluated using two main metrics, namely Accuracy and F1-macro, which are used to represent global performance and prediction balance between classes in unbalanced dataset conditions (Farhadpour et al., 2024). Hyperparameter tuning was intentionally not performed in this study, as the primary objective is to evaluate the impact of SMOTE on classification performance across different model architectures under standardised conditions, rather than to optimise individual model performance.

- Accuracy

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + T} \quad (1)$$

Accuracy measures how well a classification model can predict correctly overall, considering both positive and negative classes (Farhadpour et al., 2024).

- F1-Macro

$$F1\ \text{Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (2)$$

The F1 score is a way to balance Precision and Recall by taking their average. It helps when it's important to avoid both false positives and false negatives. (Farhadpour et al., 2024).

- Recall (Sensitivity)

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

Recall shows how well a model can find all the positive cases in the data, meaning it measures the model's ability to spot positive outcomes (Farhadpour et al., 2024).

- Precision

$$\text{Precision} = \frac{TP}{TP + FP} \quad (4)$$

Precision measures the extent to which the model can correctly predict positive outcomes. That is, how many of the positive predictions are actually positive (Farhadpour et al., 2024).

Although accuracy is still reported as a global performance indicator, the evaluation focuses on the macro F1-score because it is more representative in class imbalance conditions by giving equal weight to each class. The training evaluation results are reported as the average value of cross-validation, while the final evaluation is performed using separate test data to assess the model's generalization ability.

RESULTS AND DISCUSSION

1. Band Gap Class Distribution and Feature Correlation Analysis.

The distribution of band gap types was analyzed to identify the level of data imbalance that could potentially affect the performance of the classification model. The visualization in Figure 3 shows a comparison of the number of Direct and Indirect class samples before and after the data cleaning process. The results show that even though duplicate data has been removed, the proportion of the minority class (Indirect) remains much smaller than the majority class (Direct).

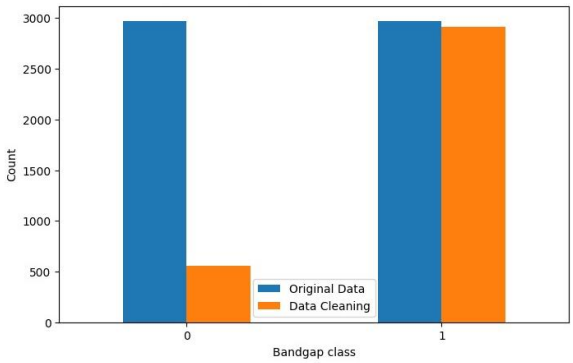

Figure 3. Class Distribution of Band Gaps Before and After Data Cleaning

This condition indicates that the data cleaning process does not significantly change the class distribution structure, so the risk of prediction bias towards the majority class remains high if no imbalance handling strategy is implemented during the model training stage.

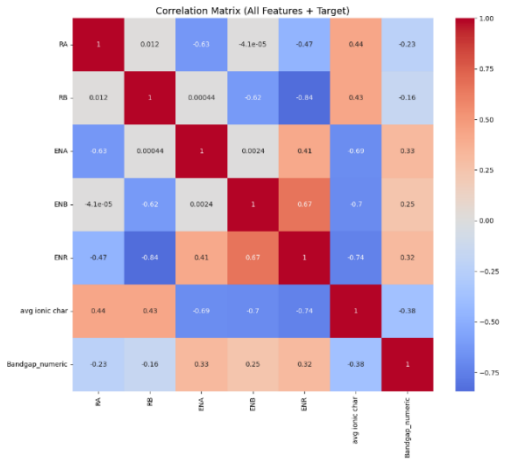

Figure 4. Pearson Correlation Matrix between Features and Band Gap Classes

The relationship between features and their correlation with band gap classes was analyzed using Pearson's correlation matrix, as shown in Figure 4. The visualization results show a strong negative correlation between ENR and RB ($r \approx -0.84$), as well as

a positive correlation between ENB and ENR ($r \approx 0.67$), indicating a close relationship between electronegativity and perovskite chemical structure. The correlations between features and target variables tend to be low to moderate, with the highest values shown by ENA and ENR, indicating that band gap type predictions are not dominated by a single feature, but rather are the result of multivariate interactions between physicochemical parameters. This pattern reinforces the justification for using non-linear and ensemble-based machine learning models to capture the complexity of the decision boundary in a non-linear feature space.

1. Classification Results in the Baseline Scenario

The results from evaluating how well the models performed in the baseline scenario, which means without adjusting for class imbalance, are shown in Table 4. Overall, all the models achieved fairly high accuracy scores during testing, between 0.87 and 0.89. However, the F1-macro scores were lower, ranging from 0.73 to 0.78, which suggests that the models are not performing equally well across all classes.

Table 4. Model Performance in the Baseline Scenario

Model	Accuracy	F1 Macro	Recall Indirect (0)	Recall Direct (1)
MLP	0.8934	0.7783	0.54	0.96
Extra-Trees	0.8703	0.7349	0.49	0.94
CatBoost	0.8847	0.7603	0.51	0.96
Gradient Bosting	0.8833	0.7563	0.50	0.96

The Multi-Layer Perceptron (MLP) model showed the highest F1-macro score in the baseline scenario, followed by CatBoost and Gradient Boosting. However, further analysis of the metrics per class shows that all models tend to have low recall in the Indirect band gap class, with values ranging from 0.49 to 0.54. In contrast, performance in the Direct band gap class is very high, with recall approaching 0.95.

These findings indicate that high accuracy in the baseline scenario is primarily driven by the model's ability to classify the majority class. The variation in model response to SMOTE generated synthetic data can be attributed to the fundamental differences in their learning mechanisms. MLP, as a neural network-based model, is highly sensitive to changes in training data distribution. The introduction of synthetic samples through SMOTE altered the feature space density around the minority class boundary, causing MLP to recalibrate its internal weights substantially. While this increased recall for the Indirect class, it simultaneously reduced precision, resulting in an overall decline in F1-macro. This behaviour is consistent with the known sensitivity of neural networks to distributional shifts in training data. In contrast, Gradient Boosting and CatBoost employ sequential boosting mechanisms that iteratively correct prediction errors from previous estimators, making them inherently adaptive to shifts in class distribution without catastrophic performance degradation. Extra Trees, on the other hand, constructs trees in parallel using random feature thresholds, which provides robustness against noise but limits its ability to adaptively respond to distribution changes resulting in modest improvement in minority class recall with relatively stable but unchanged F1-macro. In the context of imbalanced data, this condition indicates that the model is not yet able to adequately capture the characteristic patterns of the minority class, even though its overall performance appears to be good.

2. Classification Results with Class Imbalance Handling (SMOTE)

The classification results after applying class imbalance handling using SMOTE are shown in Table 5. Compared to the baseline scenario, there is a consistent change in performance patterns across all models. In general, the test accuracy value has decreased, but this is accompanied by an increase in the model's ability to recognize minority classes.

Table 5. Model Performance with Imbalance Handling (SMOTE)

Model	Accuracy	F1 Macro	Recall Indirect (0)	Recall Direct (1)
MLP	0.7752	0.6775	0.70	0.79
Extra-Trees	0.8487	0.7329	0.59	0.90
CatBoost	0.8573	0.7464	0.61	0.90
Gradient Bosting	0.8473	0.7492	0.69	0.88

The most significant improvement was seen in the MLP model, where the recall value for the Indirect band gap class increased substantially. However, this increase in sensitivity was accompanied by a decrease in the precision value for the same class, resulting in an overall decrease in the F1-macro value. This pattern shows that MLP is very sensitive to changes in the training data distribution due to synthetic samples resulting from oversampling.

In contrast, the CatBoost and Gradient Boosting models showed more stable performance. Both models were able to maintain relatively high F1-macro values with moderate increases in minority class recall without a drastic decrease in accuracy. These results indicate that boosting-based ensemble models are more robust to changes in data distribution due to the application of SMOTE.

3. Comparison of Baseline and SMOTE Performance

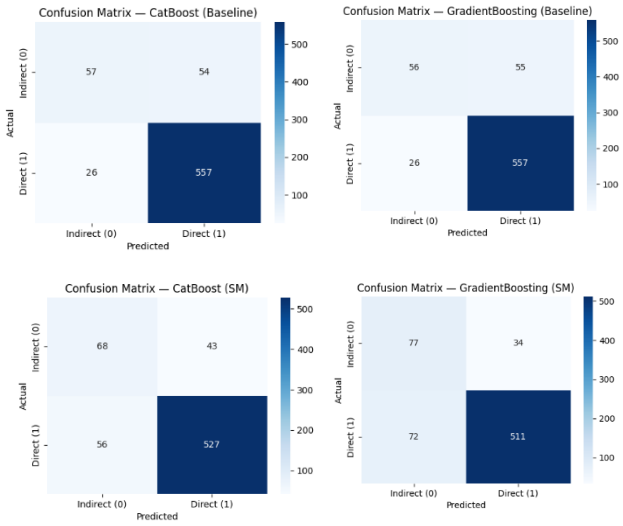

Figure 5. Comparison of model performance between baseline and SMOTE scenarios in terms of accuracy and F1-macro.

A direct comparison between the baseline scenario and SMOTE, as illustrated in Figure 5, shows a clear trade off between global accuracy and sensitivity to minority classes. In all models, the application of SMOTE did not result in an increase in accuracy, but it consistently improved the model's ability to detect the Indirect band gap class. The performance gap between accuracy and F1-macro observed across all models reinforces the argument that accuracy alone is insufficient as an evaluation metric under class imbalance conditions. The narrowing of this gap in Gradient Boosting and CatBoost after SMOTE application indicates that these models achieved a more balanced prediction distribution between the Direct and Indirect classes, whereas the widening gap in MLP suggests a precision-recall trade-off that disproportionately affected its macro-level performance.

A closer examination of the confusion matrices reveals that the improvement in Indirect class recall after SMOTE comes at

the cost of a moderate increase in false positives for the Direct class. This trade-off is most pronounced in the MLP model, where the substantial increase in Indirect recall was accompanied by a notable rise in misclassified Direct samples. Conversely, Gradient Boosting maintained a more proportional error distribution across both classes, suggesting that its boosting mechanism effectively balanced sensitivity gains without substantially compromising specificity for the majority class.

These results confirm that the application of SMOTE in this study is not intended to maximize accuracy, but rather to improve prediction fairness between classes in unbalanced datasets.

4. Model Stability Based on Cross-Validation

The results of the evaluation using 5-Fold Stratified Cross-Validation are presented in **Table 6**.

Table 6. 5-Fold Stratified Cross-Validation Results

Model	CV Accuracy (mean ± std)	CV F1-macro (mean ± std)
MLP	0.8132 ± 0.011	0.7105 ± 0.0118
Extra-Trees	0.8562 ± 0.0062	0.7329 ± 0.0130
CatBoost	0.8674 ± 0.0109	0.7522 ± 0.0236
Gradient Bosting	0.8536 ± 0.0038	0.7522 ± 0.0079

In the baseline scenario, MLP showed the highest average F1-macro score, followed by Gradient Boosting and CatBoost. The relatively small variance across all models indicates that the performance obtained is stable and does not depend on a particular data split.

In the SMOTE scenario, there was a change in model ranking based on F1-macro. CatBoost and Gradient Boosting ranked highest with the highest average F1-macro values, while MLP showed a significant decline in performance. These findings indicate that the application of SMOTE not only affects performance values, but also changes the characteristics of the model that is most suitable for classifying imbalanced data.

5. Discussion of the Impact of Class Imbalance.

The results in the baseline scenario show that class imbalance causes machine learning models to tend to form decision boundaries that are biased toward the majority class. This phenomenon is consistent with findings in previous studies, which show that models with high accuracy on imbalanced data do not necessarily have good ability to recognize minority classes. This finding is particularly evident when compared to the work of (Sudha et al., 2022), who reported an overall accuracy of approximately 91% using a Random Forest classifier on the same ABO_3 perovskite band gap dataset. While their accuracy appears superior, their evaluation relied solely on this metric without explicitly reporting per-class performance indicators such as recall or F1-macro. Given the significant class imbalance in the dataset with the Direct class comprising approximately 84% of samples a high accuracy value may predominantly reflect the model's ability to correctly classify the majority class rather than its capacity to reliably identify indirect band gap materials. In contrast, the present study demonstrates that when F1-macro is adopted as the primary evaluation metric, a more nuanced picture of model performance emerges, one that better reflects the model's ability to handle both classes equitably. This

highlights the methodological importance of metric selection in materials informatics tasks involving imbalanced datasets.

In the context of materials informatics, this condition is crucial because minority classes, such as materials with indirect band gaps, still have functional relevance in certain applications. Therefore, accuracy-based evaluation alone risks producing models that are less reliable as tools for materials exploration.

6. The Effect of SMOTE on Model Behavior: An MLP Case Study

The MLP model showed the most extreme behavioral change after SMOTE application. A significant increase in the recall of the Indirect band gap class indicates that the model became more sensitive to minority class patterns. However, the accompanying decrease in precision indicates an increase in prediction errors in that class.

This phenomenon reflects the characteristics of neural networks that are highly responsive to changes in training data distribution, especially when synthetic data is used to enrich the representation of minority classes. This behaviour is consistent with the known characteristics of neural network models, which are highly responsive to distributional shifts in training data, as their weight optimisation process is directly influenced by the density and distribution of training samples (Joloudari et al., 2023).

7. Comparative Discussion Between Models

Compared to MLP, boosting-based ensemble models, particularly CatBoost and Gradient Boosting, demonstrate better adaptability to SMOTE-generated synthetic data. Both models are able to improve sensitivity to minority classes without significantly compromising performance stability. In contrast, Extra Trees tend to overfit on training data and show limited performance improvement after SMOTE application.

These findings indicate that algorithm selection plays an important role in the success of data imbalance handling strategies, and not all models respond to oversampling in the same way.

8. Implications for Band Gap Material Predictions

These findings demonstrate that relying solely on accuracy is insufficient for evaluating model performance in band gap classification tasks involving imbalanced datasets. Adopting F1-macro as the primary evaluation metric provides a more representative assessment of model capability, particularly in capturing the minority class. This approach contributes to the development of more reliable predictive models for identifying materials with specific electronic properties in data-driven materials exploration.

This study has several limitations, including the use of synthetic data, which may not fully represent the true physicochemical distribution of real materials. Further research could explore other approaches, such as cost-sensitive learning, adaptive oversampling methods, and the addition of physicochemical features to improve prediction accuracy.

CONCLUSION

This study investigated the impact of class imbalance on the performance of machine learning models for classifying band gap types in ABO_3 perovskite oxide materials. The results demonstrate that, under the baseline setting, all models achieved relatively high overall accuracy while exhibiting systematic bias toward the majority class. As a result, the models

showed limited ability to reliably identify indirect band gap materials. These findings confirm that accuracy alone is insufficient for evaluating imbalanced datasets and that F1 macro provides a more representative assessment of performance across both classes.

The application of the Synthetic Minority Over sampling Technique (SMOTE) consistently improved model sensitivity toward the minority class, as indicated by increased recall for the indirect band gap category across all evaluated models. Among the algorithms examined, the boosting based ensemble models, particularly CatBoost and Gradient Boosting, demonstrated the most stable performance under SMOTE conditions by maintaining balanced F1 macro values without substantial reductions in overall accuracy. In contrast, the Multilayer Perceptron (MLP) model exhibited greater sensitivity to the distributional changes introduced by SMOTE, while Extra Trees showed indications of overfitting. These findings highlight that effective model selection should be aligned with both the imbalance handling strategy and the intended evaluation objectives. Furthermore, the combination of F1 macro oriented evaluation and ensemble based learning provides a more reliable framework for imbalanced material classification tasks.

The results of this study also carry important practical implications for predictive modeling in materials informatics, where accurate identification of minority class materials is essential for data driven material exploration and screening. Future research may further improve the generalizability and robustness of band gap classification models through the integration of adaptive imbalance handling approaches, such as cost sensitive learning and hybrid resampling techniques, combined with richer physicochemical descriptors and external validation using independent perovskite datasets.

REFERENCES

Akrom, M., Rustad, S., & Dipojono, H. K. (2024). A machine learning approach to predict the efficiency of corrosion inhibition by natural product-based organic inhibitors. *Physica Scripta*, *99*(3), 36006. https://doi.org/10.1088/1402-4896/ad28a9

Assegie, T. A., Elanangai, V., Paulraj, J. S., Velmurugan, M., & Devesan, D. F. (2023). Evaluation of feature scaling for improving the performance of supervised learning methods. *Bulletin of Electrical Engineering and Informatics*, *12*(3), 1833–1838. https://doi.org/10.11591/eei.v12i3.5170

Buda, M., Maki, A., & Mazurowski, M. A. (2018). *A systematic study of the class imbalance problem in convolutional neural networks*. https://doi.org/10.1016/j.neunet.2018.07.011

Chawla, N. V, Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Dawa, T., & Sajjadi, B. (2024). Exploring the potential of perovskite structures for chemical looping technology: A state-of-the-art review. *Fuel Processing Technology*, *253*, 108022. https://doi.org/10.1016/j.fuproc.2023.108022

Demircioğlu, A. (2024). Applying oversampling before cross-validation will lead to high bias in radiomics. *Scientific Reports*, *14*(1), 11563. https://doi.org/10.1038/s41598-024-62585-z

Elreedy, D., Atiya, A. F., & Kamalov, F. (2024). A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning. *Machine Learning*, *113*(7), 4903–4923. https://doi.org/10.1007/s10994-022-06296-4

Farhadpour, S., Warner, T. A., & Maxwell, A. E. (2024). Selecting and Interpreting Multiclass Loss and Accuracy Assessment Metrics for Classifications with Class Imbalance: Guidance and Best Practices. *Remote Sensing*, *16*(3), 533. https://doi.org/10.3390/rs16030533

Hoye, R. L. Z., Hidalgo, J., Jagt, R. A., Correa-Baena, J., Fix, T., & MacManus-Driscoll, J. L. (2022). The Role of Dimensionality on the Optoelectronic Properties of Oxide and Halide Perovskites, and their Halide Derivatives. *Advanced Energy Materials*, *12*(4). https://doi.org/10.1002/aenm.202100499

Joloudari, J. H., Marefat, A., Nematollahi, M. A., Oyelere, S. S., & Hussain, S. (2023). Effective Class-Imbalance Learning Based on SMOTE and Convolutional Neural Networks. *Applied Sciences*, *13*(6), 4006. https://doi.org/10.3390/app13064006

Kim, B., & Kim, J. (2020). Adjusting Decision Boundary for Class Imbalanced Learning. *IEEE Access*, *8*, 81674–81685. https://doi.org/10.1109/ACCESS.2020.2991231

Kim, D., Oh, L. S., Park, J. H., Kim, H. J., Lee, S., & Lim, E. (2022). Perovskite-based electrocatalysts for oxygen evolution reaction in alkaline media: A mini review. *Frontiers in Chemistry*, *10*. https://doi.org/10.3389/fchem.2022.1024865

Liu, D., Zhou, P., Bai, H., Ai, H., Du, X., Chen, M., Liu, D., Ip, W. F., Lo, K. H., Kwok, C. T., Chen, S., Wang, S., Xing, G., Wang, X., & Pan, H. (2021). Development of Perovskite Oxide-Based Electrocatalysts for Oxygen Evolution Reaction. *Small*, *17*(43). https://doi.org/10.1002/smll.202101605

Mujahid, M., Kına, E., Rustam, F., Villar, M. G., Alvarado, E. S., Diez, I. D. L. T., & Ashraf, I. (2024). Data oversampling and imbalanced datasets: an investigation of performance for machine learning and feature engineering. *Journal of Big Data*, *11*(1), 87. https://doi.org/10.1186/s40537-024-00943-4

Mukherjee, M., & Khushi, M. (2021). SMOTE-ENC: A Novel SMOTE-Based Method to Generate Synthetic Data for Nominal and Continuous Features. *Applied System Innovation*, *4*(1), 18. https://doi.org/10.3390/asi4010018

Owusu-Adjei, M., Hayfron-Acquah, J. Ben, Frimpong, T., & Abdul-Salaam, G. (2023). Imbalanced class distribution and performance evaluation metrics: A systematic review of prediction accuracy for determining model performance in healthcare systems. *PLOS Digital Health*, *2*(11), e0000290. https://doi.org/10.1371/journal.pdig.0000290

Park, H.-J., Koo, Y.-S., Yang, H.-Y., Han, Y.-S., & Nam, C.-S. (2024). Study on Data Preprocessing for Machine Learning Based on Semiconductor Manufacturing Processes. *Sensors*, *24*(17), 5461. https://doi.org/10.3390/s24175461

Rahman, I. F., Azies, H. Al, & Akrom, M. (2025). Deteksi Struktur Material Perovskit ABO3 Berbasis Machine Learning. *Jurnal Pendidikan Fisika Dan Teknologi*, *9*(1), 2025. https://doi.org/10.47002/metik.v9i1.1036

Ramadhan, N. G. (2025). Enhancing SMOTE Using Euclidean Weighting for Imbalanced Classification Dataset. *Journal of Applied Data Sciences*, *6*(3), 2207–2220. https://doi.org/10.47738/jads.v6i3.798

Shi, J., Zhang, J., Yang, L., Qu, M., Qi, D., & Zhang, K. H. L. (2021). Wide Bandgap Oxide Semiconductors: from Materials Physics to Optoelectronic Devices. *Advanced Materials*, *33*(50). https://doi.org/10.1002/adma.202006230

Sudha, P. G., Mattur, M. N., Nagappan, N., Rath, S., & Thomas, T. (2022). Prediction of nature of band gap of perovskite oxides (ABO 3) using a machine learning approach. *Journal of Materiomics*, *8*(5), 937–948. https://doi.org/10.1016/j.jmat.2022.04.006

Szeghalmy, S., & Fazekas, A. (2023). A Comparative Study of the Use of Stratified Cross-Validation and Distribution-Balanced Stratified Cross-Validation in Imbalanced Learning. *Sensors*, *23*(4), 2333. https://doi.org/10.3390/s23042333

Wang, T., Tan, X., Wei, Y., & Jin, H. (2021). Accurate bandgap predictions of solids assisted by machine learning. *Materials Today Communications*, *29*, 102932. https://doi.org/10.1016/j.mtcomm.2021.102932

Xu, P., Ji, X., Li, M., & Lu, W. (2023). Small data machine learning in materials science. *Npj Computational Materials*, *9*(1), 42. https://doi.org/10.1038/s41524-023-01000-z

Zhang, J., Li, Y., & Zhou, X. (2023). Machine-Learning Prediction of the Computed Band Gaps of Double Perovskite Materials. *Computer Science and Machine Learning Trends 2023*, 15–27. https://doi.org/10.5121/csit.2023.130102

Zhao, J., Wang, X., Li, H., & Xu, X. (2024). Interpretable machine learning-assisted screening of perovskite oxides. *RSC Advances*, *14*(6), 3909–3922. https://doi.org/10.1039/D3RA08591K