

Comparative Analysis of SMOTE-Tomek and SMOTE-ENN for Machine Learning Based Band Gap Classification of ABO₃ Perovskite Oxides

Chalida Abdat^{1*}, Muhamad Akrom^{2**}, Johana Oktavia Ramadhani^{3*}, De Rosal Ignatius Moses Setiadi^{4**}

^{*}Informatics Engineering, Faculty of Computer Science, Universitas Dian Nuswantoro

^{**}Research Center for Quantum Computing and Materials Informatics, Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang 50131 Indonesia

111202315031@mhs.dinus.ac.id¹, m.akrom@dsn.dinus.ac.id², 111202315024@mhs.dinus.ac.id³, moses@dsn.dinus.ac.id⁴

Article Info

Article history:

Received 2026-07-27

Revised 2026-09-11

Accepted 2026-10-05

Keyword:

Band Gap,
Class Imbalance,
Perovskite Oxide,
SMOTE-ENN,
SMOTE-Tomek.

ABSTRACT

The electronic band gap of ABO₃ perovskite oxides is a key property for photovoltaic and optoelectronic applications, where distinguishing direct from indirect band gaps guides material selection. Determining band gap type through first principles calculations or experiments is accurate but computationally expensive and time consuming, motivating machine learning as a faster alternative. A practical obstacle in this setting is class imbalance: Direct band gap samples substantially outnumber indirect ones in the dataset under study, which pushes classifiers toward majority class bias. This study examines how two hybrid resampling techniques, SMOTE-Tomek and SMOTE-ENN, affect the classification of band gap type in ABO₃ perovskites using six physicochemical descriptors. Four classifiers (Multilayer Perceptron, Extra Trees, Gradient Boosting, and CatBoost) were trained on an 80:20 stratified split with fivefold cross validation, and the best performing model was further tuned with Optuna. Resampling was applied only to the training data, and models were evaluated on accuracy, precision, recall, and macro F1 score using a held out test set. CatBoost proved the most consistent model under both resampling schemes. SMOTE-Tomek produced a more balanced training distribution and higher test performance (accuracy 0.86, macro F1 0.76) than SMOTE-ENN (accuracy 0.81, macro F1 0.72). Hyperparameter tuning further improved CatBoost's cross validated macro F1 score. The results indicate that resampling mainly improves minority class representation rather than overall accuracy, and that electronegativity related descriptors are the most informative features for predicting band gap type.



This is an open access article under the [CC-BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.

I. INTRODUCTION

ABO₃ perovskite oxides are a class of functional materials with attractive structural and electronic characteristics for a variety of energy and optoelectronic applications, including solar cells, light-emitting devices, and photocatalysts [1], [2], [3].

One important electronic characteristic used to evaluate the potential of these materials is the band gap, including whether it is classified as a direct or indirect band gap. In a direct band gap, electronic transitions between the valence band and the

conduction band can occur without a significant change in crystal momentum, whereas in an indirect band gap the transition involves a momentum change that is generally associated with phonon interactions [4], [5]. The difference in these transition mechanisms makes the band gap type an important parameter in material characterization and candidate screening. As the number of material candidates to be evaluated increases, an approach capable of classifying band gap type efficiently is needed to support the early-stage materials screening process.

Conventionally, band gap characteristics can be determined through Density Functional Theory (DFT) calculations or experimental characterization. These approaches provide important information about a material's structure and electronic properties, but evaluating a large number of candidates can require significant time and computational or experimental resources [6], [7]. In the context of materials informatics, machine learning (ML) can be used as a complementary approach to build predictive models based on descriptors that represent the chemical and physical characteristics of materials [1]. Such models can be used to perform rapid screening of candidate materials before further analysis using first-principles calculations or more expensive experimental validation. Thus, the application of ML has the potential to accelerate the screening of ABO_3 candidates based on band gap type without eliminating the role of computational and experimental methods as a validation stage.

However, applying machine learning to material property classification can face challenges when the class distribution is imbalanced. The ABO_3 dataset used in this study consists of approximately 84% direct band gap samples and 16% indirect band gap samples, with an imbalance ratio of about 5.2:1. This distribution can cause the model to prioritize the majority class, making its ability to recognize the minority class less optimal. In such conditions, accuracy alone is not sufficient to represent classification performance in a balanced way, because a high accuracy value can still be achieved even when the model's ability to recognize the minority class is relatively limited [8], [9]. Therefore, evaluation on imbalanced datasets needs to take the performance of each class into account, for example through the macro F1-score, which gives equal weight to each class.

One approach that can be used to address class imbalance is resampling, particularly through a combination of minority sample generation and data cleaning. Synthetic Minority Over-sampling Technique (SMOTE) generates synthetic samples for the minority class based on relationships among minority samples [8], [9]. Although this approach improves the representation of the minority class, adding synthetic samples in regions of the data where classes overlap can alter the local structure of the data and the position of the decision boundary learned by the model. Therefore, hybrid resampling combines SMOTE with data-cleaning methods to reduce samples that could potentially interfere with the formation of the decision boundary. SMOTE-Tomek combines SMOTE with the removal of Tomek links, while SMOTE-ENN combines SMOTE with Edited Nearest Neighbors (ENN) [8], [10]. Both methods aim to address class imbalance, but they use different cleaning mechanisms, which can lead to different changes in the data's composition, representation, and local structure [11], [12].

The difference in these mechanisms shows that the effectiveness of resampling cannot be evaluated solely on how balanced the resulting class distribution is. Resampling

can alter the data representation and local structure the model uses to form the decision boundary, so its impact on performance may depend on the characteristics of the learning algorithm used. Each classifier has a different mechanism for leveraging data patterns and structure, so distribution changes caused by resampling can produce different performance responses across models. Thus, improved class balance cannot be directly assumed to lead to better classification performance and generalization. Evaluation of resampling needs to consider how the effects of resampling strategies differ across classifiers and compare them with the no-resampling condition as a baseline to determine whether the resulting changes in data distribution truly provide predictive benefit.

A number of studies have shown the application of ML to predict perovskite properties and band gaps using various descriptors and algorithms [2], [3], [6], [7][3], [13]. On the other hand, research on handling class imbalance has evaluated various resampling strategies, including SMOTE-Tomek and SMOTE-ENN across different classification problems [11], [12]. However, based on the literature reviewed in this study, no study has yet been found that systematically compares SMOTE-Tomek and SMOTE-ENN head-to-head for classifying band gap types in ABO_3 using a uniform experimental protocol across multiple classifiers at once. In addition, the consistency of the impact of resampling strategies on classifiers with different learning mechanisms has not been widely evaluated in the context of ABO_3 band gap type classification. This is important because resampling not only changes class proportions, but can also alter the local data representation and decision boundary used by the classifier. A previous study by some of the authors of this research applied a single Tomek-Link approach (without SMOTE oversampling) to ABO_3 band gap classification using one classifier [14]. This study differs from those studies in several respects: (1) two hybrid resampling strategies were evaluated comparatively SMOTE-Tomek and SMOTE-ENN against a no-resampling baseline; (2) the evaluation was conducted on four classifiers with different learning characteristics to test whether the impact of resampling is consistent across classifiers; and (3) the best configuration then underwent systematic hyperparameter optimization and feature importance analysis, which were not performed in the previous studies. Therefore, a consistent experimental framework is needed to evaluate the effect of resampling strategies on classification performance and classifier responses to changes in data representation.

Based on these research problems and gaps, this study evaluates the effect of resampling strategies on direct and indirect band gap classification in ABO_3 perovskite oxides. The experiments compare three data-processing conditions: no resampling as the baseline, SMOTE-Tomek, and SMOTE-ENN, using four classifiers with different learning characteristics: Multi-Layer Perceptron (MLP), Extra Trees, Gradient Boosting, and CatBoost. The comparison is carried out using a stratified train-test split with five-fold cross-

validation on the training data, while macro F1-score is used as the main metric and other evaluation metrics as supporting measures. Specifically, this study aims to: (1) evaluate baseline performance under imbalanced class conditions; (2) compare the impact of SMOTE-Tomek and SMOTE-ENN on classification performance, including whether improved class balance through resampling consistently improves performance over the baseline or instead creates a trade-off between sensitivity to the minority class and overall classification performance; and (3) analyze whether the effects of the two resampling strategies differ across classifiers.

To complement the performance evaluation, the selected model is analyzed using feature importance to identify the physicochemical descriptors that contribute most strongly to band gap type prediction. Thus, the analysis focuses not only on changes in classification metrics due to resampling, but also on the interpretation of descriptors in distinguishing direct and indirect band gaps. This study provides a systematic evaluation of the effect of SMOTE-Tomek and SMOTE-ENN on ABO₃ band gap type classification by comparing several classifiers within a consistent experimental protocol. The evaluation focuses on differences in classifier responses to resampling strategies and their implications for classification performance, so that this study can provide a clearer understanding of whether resampling delivers consistent benefits across different learning mechanisms.

II. METHODS

This study uses a supervised machine learning approach to classify band gap types in ABO₃ perovskite oxide materials. The classification is based on six physicochemical descriptors. The research stages include data exploration, data preprocessing, class imbalance handling, resampling, model training, and performance evaluation. The overall research workflow is shown in Figure 1.

A. Dataset Description

The dataset used in this study refers to the dataset collected by Priyanga et al. [15] titled “*Prediction of nature of band gap of perovskite oxides (ABO₃) using a machine learning approach*”. This dataset contains ABO₃ perovskite materials along with their band gap type classifications. The original dataset consisted of 5,940 samples with seven attributes, including six predictor attributes and one target label: A-site ionic radius (RA), B-site ionic radius (RB), A-site electronegativity (ENA), B-site electronegativity (ENB), electronegativity ratio (ENR), average ionic character (avg. ionic char), and band gap type [15].

Six descriptors are used as predictor variables to represent the physicochemical characteristics of ABO₃ materials. RA and RB represent the ionic radius at the A and B sites, which are related to the size of the constituent ions and the material’s structural characteristics. ENA and ENB represent the electronegativity at each site, which describes an element’s

tendency to attract electrons. ENR represents the electronegativity ratio between the A and B sites, while avg. ionic char represents the material’s average ionic character. These six descriptors are used as a numerical representation of the chemical and structural characteristics of the ABO₃ constituents in the band gap type classification process. The definitions and formulations of the descriptors used follow the source dataset [15].

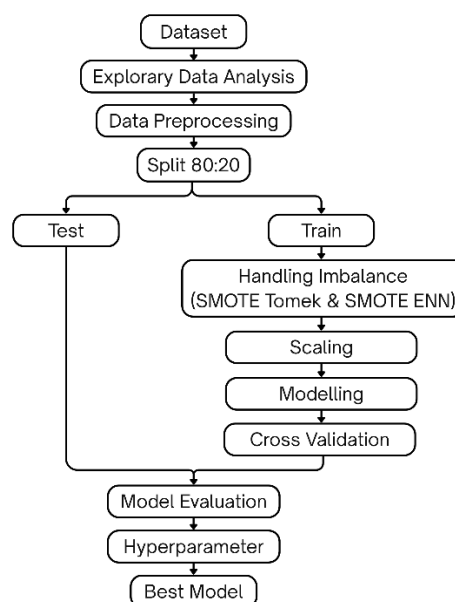


Figure 1. Research Flowchart

Before the classification experiments were conducted, duplicate records were removed based on identical values across all predictor attributes and the target label, retaining one record for each identical entry. A total of 2,471 duplicate records were removed from the original dataset, resulting in 3,469 unique samples. Inspection of the cleaned dataset showed that there were no missing values. The target label was encoded into two classes: 1 for direct band gap and 0 for indirect band gap. After cleaning, the direct band gap class consisted of 2,913 samples (84%), while the indirect band gap class consisted of 556 samples (16%), with a class imbalance ratio of approximately 5.2:1. A representative excerpt of the dataset is shown in Table I.

TABLE I
EXCERPT OF THE ABO₃ PEROVSKITE OXIDE DATASET

RA	RB	ENA	ENB	ENR	avg ionic char	Band gap
1.05	1.18	1.25	0.95	-3.306393	0.179242	0
1.12	0.27	1.10	2.04	-1.117357	0.143887	0
0.80	0.71	1.66	1.30	-1.935072	0.148737	0
0.83	0.78	1.30	1.83	-2.031679	0.141753	0
1.08	0.65	1.22	1.55	-2.042928	0.156943	0
...	...	...	...	...	...	...

B. Exploratory Data Analysis (EDA) and Data Preprocessing

Exploratory Data Analysis (EDA) was conducted to understand the characteristics of the dataset and inspect the data conditions before the classification process. The analysis included examining the class distribution, the distribution of each descriptor, and the relationships among descriptors. The distributions of the six descriptors, namely RA, RB, ENA, ENB, ENR, and avg. ionic char, were visualized using histograms with kernel density estimation (KDE), as shown in Figure 2. Correlation analysis among descriptors was then performed using a correlation matrix, as shown in Figure 3.

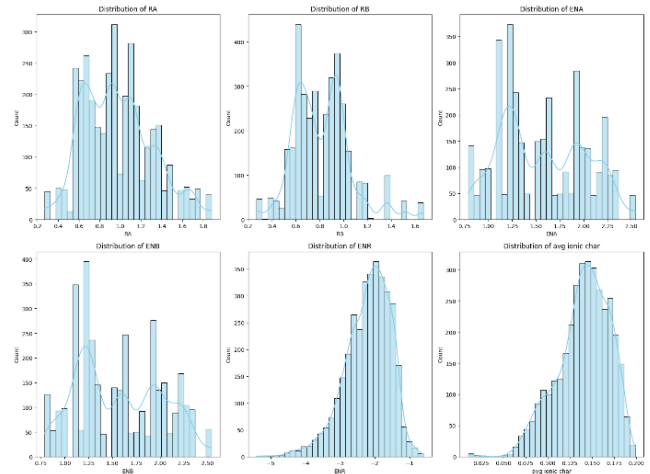
The descriptor distributions showed variation and different distribution patterns across variables, with some descriptors exhibiting distributions that were not entirely unimodal. The correlation analysis results indicated relatively strong relationships for several descriptor pairs, especially between ENA and avg. ionic char and between RA and RB. This information was used to understand the characteristics and interrelationships among descriptors before the modeling process. The variation in distributions and the relationships among descriptors also provided important context for evaluating how resampling strategies could affect the data representation used by the classifier.

After the data inspection and cleaning process, the dataset was split into an 80% training set and a 20% test set using stratified random sampling with `random_state = 42`. Stratification was performed to preserve the class proportions in both subsets. Of the 3,469 samples, 2,775 samples were used as the training set and 694 samples as the test set. The class distribution in both subsets before the resampling process is shown in Table II.

TABLE II

CLASS DISTRIBUTION OF TRAINING AND TEST DATA (BEFORE RESAMPLING)

Subset	Direct (1)	Indirect (0)	Total
Train	2,330	445	2,775
Test	583	111	694

Figure 2. Distribution of the physicochemical descriptors in the ABO₃

perovskite oxide dataset.

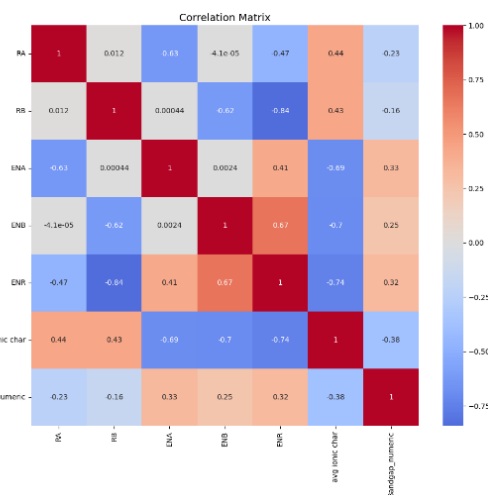


Figure 3. Correlation matrix of the physicochemical descriptors

C. Class Imbalance and Resampling Strategy

The dataset used in this study has an imbalanced class distribution, with direct band gap as the majority class and indirect band gap as the minority class. In the training set, there are 2,330 direct band gap samples and 445 indirect band gap samples. This results in a class ratio of about 5.2:1 and serves as the basis for evaluating resampling strategies in the model training process.

The no-resampling condition is used as the baseline to measure model performance on the original data distribution. The baseline performance is then compared with the performance of the model after applying two hybrid resampling strategies, namely SMOTE-Tomek and SMOTE-ENN. In this way, the effect of resampling can be evaluated relative to the condition without class distribution modification.

SMOTE-Tomek and SMOTE-ENN both combine minority class oversampling using Synthetic Minority Over-sampling Technique (SMOTE) with a data-cleaning mechanism. The main difference between the two strategies lies in the cleaning method used. SMOTE-Tomek uses Tomek Links, while SMOTE-ENN uses Edited Nearest Neighbors (ENN) [8], [9], [10], [12], [10], [16].

Resampling is applied only to the training data, while the test set is not resampled and is used as a held-out test set to evaluate the model's generalization ability. In the five-fold cross-validation process performed on the training set, resampling is applied separately within each training fold, while the validation fold continues to use the original data distribution. With this procedure, synthetic samples generated during resampling are not used in the validation process, so the risk of data leakage across folds can be minimized.

SMOTE-Tomek is a hybrid resampling method that combines SMOTE with Tomek Links [8], [9], [10], [16]. SMOTE improves minority-class representation by generating synthetic samples based on proximity relationships among minority samples in feature space [8], [9]. After that, Tomek Links are used to identify pairs of samples from different classes that are highly similar and lie in adjacent regions near the inter-class boundary [10], [16]. This mechanism is used to remove samples that could increase ambiguity at the decision boundary, so the structure of the resampled data may differ from that produced by SMOTE without a cleaning step.

SMOTE-ENN combines SMOTE with Edited Nearest Neighbors (ENN) [8], [9], [10], [12]. SMOTE is used to improve minority-class representation through the creation of synthetic samples, while ENN performs data cleaning based on the consistency of a sample's label with the labels of its nearest neighbors [10], [12]. Samples deemed inconsistent under the nearest-neighbor rule may be removed during the ENN process. This cleaning mechanism, which differs from Tomek Links, can produce a different composition and local structure in the training data, potentially resulting in different performance responses across classifiers.

D. Classification Models

This study uses four classification algorithms: Multi-Layer Perceptron (MLP), Extra Trees, Gradient Boosting, and CatBoost. The four classifiers were selected to represent different learning characteristics, making it possible to evaluate whether the effect of resampling strategies is consistent or depends on the classifier's learning mechanism.

Multi-Layer Perceptron (MLP) is a neural network-based model that learns nonlinear relationships between descriptors and the target through hidden layers. In this study, MLP uses the Adam optimization algorithm with `max_iter = 500`. Feature standardization is applied to MLP because neural network-based models are sensitive to differences in scale across features. The standardization parameters are determined from the training data and applied to the

validation and test data to maintain consistency in the evaluation process.

Extra Trees is an ensemble method based on decision trees that introduces a high degree of randomness in the tree-building process [17]. Randomness is applied in selecting splits for each tree, and the predictions from all trees are then combined to produce the final prediction [17]. Extra Trees is used to represent a tree-ensemble approach with a high level of randomness.

Gradient Boosting is an ensemble method that builds a number of decision trees sequentially, with each subsequent model aimed at correcting the prediction errors of the previous model [18]. This approach is used as a representation of tree-based boosting methods in the experiment.

CatBoost is a decision tree-based gradient boosting algorithm that uses an ordered boosting mechanism in the learning process [19]. CatBoost is used as a boosting-based classifier with a learning mechanism different from Gradient Boosting [19].

Thus, MLP represents nonlinear neural network-based learning, Extra Trees represents a tree ensemble with high randomness, while Gradient Boosting and CatBoost represent tree-based gradient boosting approaches with different learning mechanisms. This selection makes it possible to evaluate the effect of resampling strategies on classifiers with diverse learning characteristics.

E. Experimental Design and Validation

The experiment was designed to compare the effects of three data-processing conditions across four classifiers. The three conditions were baseline without resampling, SMOTE-Tomek, and SMOTE-ENN, while the classifiers used were Multi-Layer Perceptron (MLP), Extra Trees, Gradient Boosting, and CatBoost. Combining the three data conditions with the four classifiers produced 12 experimental configurations, all evaluated using the same experimental protocol.

Each configuration was evaluated using 5-fold stratified cross-validation on the training set to preserve class proportions across folds. In each fold, preprocessing parameters such as feature standardization were fitted exclusively on the training fold before transforming both training and validation folds. For SMOTE-Tomek and SMOTE-ENN conditions, resampling was applied solely to the preprocessed training fold, leaving the validation fold unresampled to evaluate performance on its original class distribution.

This procedure was applied independently to each combination of classifier and resampling condition, ensuring no synthetic samples entered validation folds and preventing data leakage during training. The performance values across the five validation folds were summarized using the mean and standard deviation to capture both average performance and fold-to-fold variability.

Macro F1-score served as the primary metric because it weights the direct and indirect band gap classes equally, making it ideal for evaluating imbalanced datasets. Accuracy, class-wise precision, class-wise recall, and ROC-AUC were used as supporting metrics to provide a comprehensive evaluation of each configuration's classification performance.

Following cross-validation, the best configuration was selected based on the highest average macro F1-score across validation folds. The held-out test set was completely excluded from hyperparameter tuning and model selection. After selecting the optimal configuration, the final model was evaluated on the unresampled test set in its original distribution to assess generalization performance on unseen data.

F. Evaluation Metrics

Model performance was evaluated using accuracy, balanced accuracy, precision, recall, macro F1-score, and ROC-AUC. Macro F1-score served as the primary metric because it gives equal weight to both direct and indirect band gap classes, making it ideal for imbalanced dataset evaluation. The remaining metrics provided supporting insights to comprehensively capture each model's performance characteristics.

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

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

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

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

Notes:

TP: True Positive

TN: True Negative

FP: False Positive

FN: False Negative

This formulation ensures that performance on the indirect band gap class is not overshadowed by the dominant direct class. ROC-AUC is used as a supplementary metric to assess class separability across various classification thresholds rather than as the primary basis for selecting model configurations.

G. Hyperparameter Optimization

Hyperparameter optimization was performed following the initial evaluation to refine the top-performing classifier configuration based on the macro F1-score. Optimization was conducted separately for each resampling strategy, tailoring hyperparameter configurations to the unique training data characteristics generated by each technique. Using Optuna with the Tree-structured Parzen Estimator (TPE) sampler [20], the search process maximized the average macro F1-

score objective over 100 trials using a fixed random seed of 42. Each trial was evaluated via 5-fold stratified cross-validation on the training set, where all preprocessing and resampling steps were strictly constrained to the training fold using a pipeline to prevent data leakage into the validation fold.

The held-out test set remained entirely untouched during both hyperparameter optimization and configuration selection. Optimization focused on relevant parameters of the selected baseline classifier, and the configuration achieving the highest mean cross-validation macro F1-score was selected to train the final model. This finalized model was subsequently evaluated on the held-out test set to provide an unbiased estimate of generalization performance on unseen data.

H. Feature Importance Analysis

Feature importance analysis was performed on the best-performing model configuration from the previous stage to identify the physicochemical descriptors making the largest predictive contribution to band gap type classification. Calculated using the model's built-in interpretation mechanism, the derived importance values were compared to evaluate the relative influence of each descriptor during training and identify the most dominant predictors.

The feature importance results served as an interpretive complement to the classification performance evaluation. Importance values were interpreted strictly as measures of relative predictive contribution to the model's decisions, rather than evidence of a causal relationship between the descriptors and band gap type. The domain interpretation of the dominant features was then conducted by considering the fundamental characteristics and physicochemical meaning of those descriptors.

III. RESULT AND DISCUSSION

A. Data Distribution after Resampling

After preprocessing, the dataset consisted of 3,469 samples, with 2,913 samples (84%) in the Direct class and 556 samples (16%) in the Indirect class. After a stratified train-test split, the training set contained 2,775 samples 2,330 Direct and 445 Indirect while 694 samples were held out as the test set. Thus, the class ratio in the training set before resampling was approximately 5.24:1, indicating a fairly strong imbalance that could cause the classifier to lean toward the majority class.

Resampling was applied only to the training set, while the held-out test set was kept in its original distribution to ensure that evaluation reflected data that had not been resampled. The two hybrid strategies produced different training populations. SMOTE-Tomek yielded 4,524 samples, consisting of 2,262 Direct and 2,262 Indirect, resulting in a balanced 50:50 distribution. SMOTE-ENN, by contrast, produced 3,545 samples, consisting of 1,856 Direct and 1,689

Indirect. These results show that although both strategies aim to reduce class imbalance through a combination of oversampling and cleaning, the cleaning process led to different degrees of sample reduction.

The difference in sample counts after resampling is important for interpreting the classification results in the next section. SMOTE-Tomek preserved a larger training population, whereas SMOTE-ENN produced a smaller training set after cleaning. As a result, the two strategies not only created class distributions that differed from the original condition, but also formed different training representations, which could in turn affect the decision boundary and the behavior of each classifier.

TABLE III
CLASS DISTRIBUTION BEFORE AND AFTER RESAMPLING

Dataset Condition	Direct (%)	Indirect (%)	Total
Training set	2,330 (84.0%)	445 (16.0%)	2,775
SMOTE-Tomek	2,262 (50.0%)	2,262 (50.0%)	4,524
SMOTE-ENN	1,856 (52.4%)	1,689 (47.6%)	3,545

Note: The held-out test set did not undergo any resampling process and was used only at the model generalization evaluation stage.

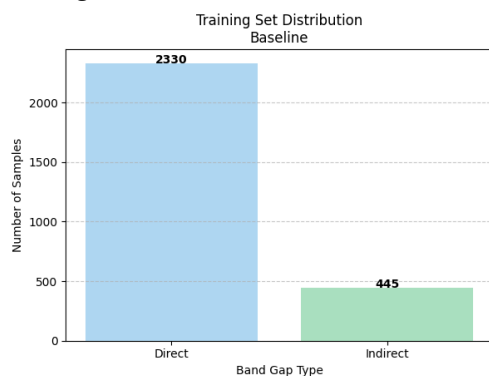


Figure 4. Training data class distribution before resampling

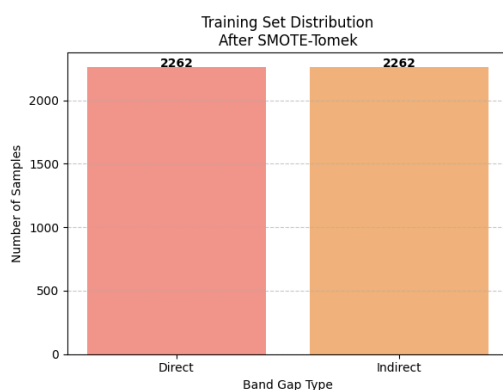


Figure 5. Training data class distribution after SMOTE-Tomek

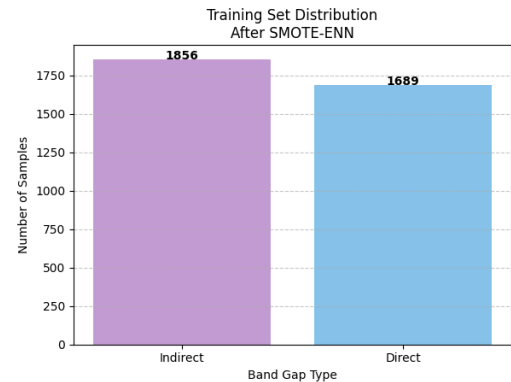


Figure 6. Training data class distribution after SMOTE-ENN

B. Baseline Performance

The baseline serves as the reference condition to evaluate classifier performance on the original imbalanced data distribution without resampling. Evaluated on the held-out test set after training on imbalanced data, these results provide a benchmark for identifying shifts in performance and model behavior following the application of SMOTE-Tomek and SMOTE-ENN. The baseline performance for each model on the held-out test set is presented in Table IV.

TABLE IV
MODEL BASELINE PERFORMANCE

Model	Accuracy	Macro F1	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 Boosting	0.8833	0.7563	0.50	0.96

Based on Table IV, baseline models achieved high overall accuracy (0.8703–0.8934), led by MLP (0.8934), CatBoost (0.8847), Gradient Boosting (0.8833), and Extra Trees (0.8703). However, these values must be interpreted alongside macro F1-scores, which ranged from 0.7349 (Extra Trees) to 0.7783 (MLP). The disparity between accuracy and class-wise performance is evident in recall values: Indirect class (0) recall ranged from only 0.49 (Extra Trees) to 0.54 (MLP), whereas Direct class (1) recall consistently reached 0.94–0.96 across all models.

This imbalance stems from the majority-dominated training set, where high overall accuracy primarily reflects successful Direct-class recognition rather than balanced predictive capacity. As confirmed by the confusion matrices in Figure 7, misclassifications occurred far more frequently for Indirect samples. Consequently, baseline evaluations demonstrate that the primary challenge lies in identifying the minority class rather than recognizing the majority class.

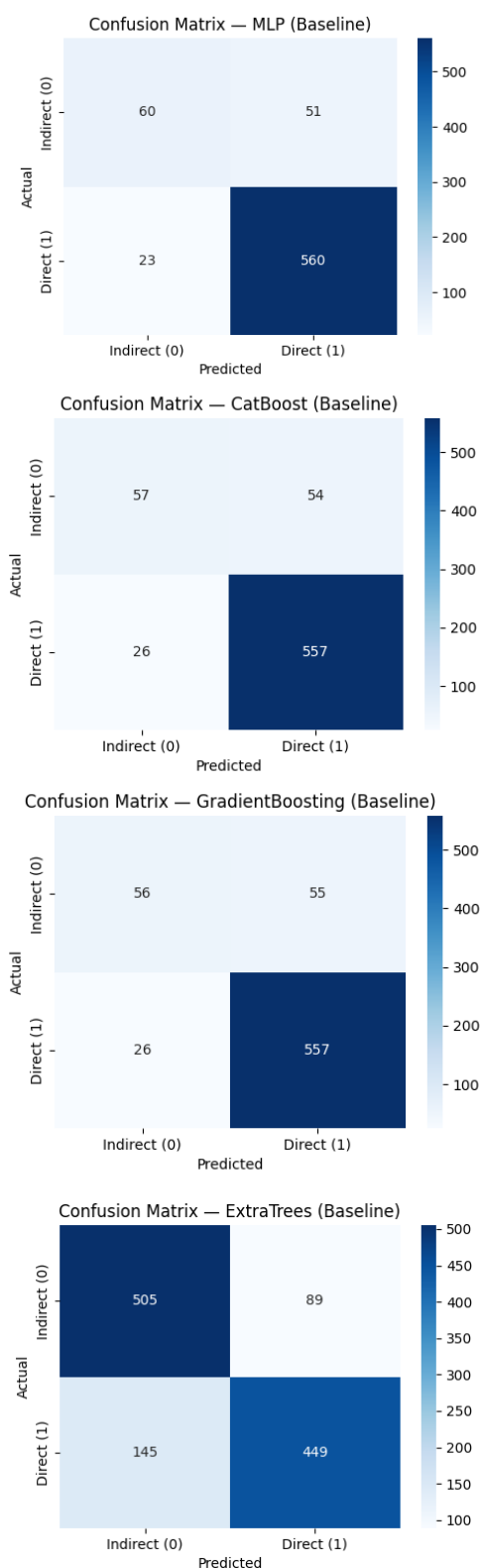


Figure 7. Test set confusion matrix for the baseline models

Thus, macro F1-score and class wise recall became important metrics for evaluating the baseline condition,

because they provide more representative information about the balance of performance across classes than accuracy alone. These baseline findings then served as a reference point for assessing whether changes to the training representation through SMOTE-Tomek and SMOTE-ENN could improve sensitivity to the Indirect class without causing an unfavorable drop in Direct class recognition.

To complement the evaluation on the held-out test set, the classifier's performance consistency on the training set was evaluated using 5-fold stratified cross-validation. Stratification was used to preserve class proportions in each fold, while no resampling was applied in the baseline. The cross-validation results are presented in Table V.

TABLE V
BASELINE CROSS-VALIDATION PERFORMANCE

Model	Accuracy (Mean $\pm$ SD)	Macro F1 (Mean $\pm$ SD)
MLP	0,8879 $\pm$ 0,0019	0,7593 $\pm$ 0,0131
CatBoost	0,8864 $\pm$ 0,0038	0,7507 $\pm$ 0,0094
Gradient Boosting	0,8890 $\pm$ 0,0032	0,7553 $\pm$ 0,0066
Extra Trees	0,8732 $\pm$ 0,0047	0,7367 $\pm$ 0,0097

Based on the macro F1-score, MLP achieved the highest average value at 0.7593 ± 0.0131 , followed by Gradient Boosting at 0.7553 ± 0.0066 , CatBoost at 0.7507 ± 0.0094 , and Extra Trees at 0.7367 ± 0.0097 . Meanwhile, based on accuracy, Gradient Boosting achieved the highest average value at 0.8890 ± 0.0032 , followed by MLP at 0.8879 ± 0.0019 , CatBoost at 0.8864 ± 0.0038 , and Extra Trees at 0.8732 ± 0.0047 .

The discrepancy in model rankings underscores how metric selection influences performance interpretation under class imbalance. While Gradient Boosting achieved the highest mean accuracy, MLP recorded the highest mean macro F1-score, confirming that these metrics capture distinct operational aspects. For this study, macro F1-score serves as the primary evaluation metric due to its equal weighting of Direct and Indirect classes.

Low standard deviations across folds for accuracy (0.0019–0.0047) and macro F1-score (0.0066–0.0131) reflect high cross-validation stability under baseline conditions. However, this consistency does not negate the observed inter-class performance gap on the held-out test set, demonstrating that cross-validation stability is distinct from minority-class sensitivity.

Overall, baseline evaluations establish two key characteristics: high accuracy with consistent cross-validation performance across all classifiers, alongside a persistent performance gap between Direct and Indirect classes on unseen data. Consequently, the baseline serves not to prove model superiority, but as a benchmark to measure performance shifts following resampling. These findings ground the subsequent evaluation of SMOTE-Tomek and SMOTE-ENN, focusing on how each strategy alters minority sensitivity, majority class retention, and overall generalization.

C. Performance Comparison of SMOTE-Tomek and SMOTE-ENN

The performance of the classifiers after applying SMOTE-Tomek and SMOTE-ENN on the held-out test set is presented in Table VI. The four classifiers were evaluated using the same experimental protocol, while resampling was applied only to the training set. The held-out test set was not resampled, so it retained the original class distribution and was used to measure the model's generalization ability.

TABLE VI
INITIAL HELD-OUT TEST PERFORMANCE ACROSS RESAMPLING TECHNIQUES

Model	SMOTE-Tomek		SMOTE-ENN	
	Accuracy	F1 Macro	Accuracy	F1 Macro
MLP	0.79	0.70	0.77	0.68
Extra Trees	0.84	0.73	0.81	0.72
CatBoost	0.86	0.76	0.81	0.72
Gradient Boosting	0.84	0.74	0.79	0.70

Based on Table VI, SMOTE-Tomek produced higher accuracy and macro F1-scores than SMOTE-ENN across all classifiers. CatBoost achieved the highest performance with SMOTE-Tomek, with an accuracy of 0.86 and a macro F1-score of 0.76, whereas the highest performance with SMOTE-ENN reached an accuracy of 0.81 and a macro F1-score of 0.72. This pattern shows that, in this study's experimental setting, SMOTE-Tomek delivered higher overall performance than SMOTE-ENN.

However, compared with the baseline in Table IV, applying resampling did not automatically improve overall performance. With SMOTE-Tomek, for example, CatBoost's accuracy and macro F1 changed from 0.8847 and 0.7603 to 0.86 and 0.76. A similar decline was also observed in the other classifiers, and it was more pronounced with SMOTE-ENN. These findings indicate that successfully balancing the class distribution in the training set does not always lead to better generalization on the held-out test set. Therefore, the effectiveness of resampling needs to be assessed together with changes in the model's ability to recognize each class.

To evaluate performance during model development, 5-fold stratified cross-validation was performed on the training set. The cross-validation results and ROC-AUC on the held-out test set are presented in Table VII.

TABLE VII
CROSS-VALIDATION PERFORMANCE AND HELD-OUT TEST ROC-AUC

Model	SMOTE-Tomek (CV) + ROC-AUC (Test)				SMOTE-ENN (CV) + ROC-AUC (Test)			
	Acc	F1	95% CFI	ROC-AUC	Acc	F1	95% CFI	ROC-AUC
CB	0.87	0.75	[0.72 - 0.77]	0.85	0.82	0.72	[0.70 - 0.73]	0.86
GB	0.85	0.74	[0.72 - 0.76]	0.85	0.81	0.72	[0.71 - 0.73]	0.85
ET	0.85	0.73	[0.70 - 0.75]	0.82	0.82	0.72	[0.70 - 0.73]	0.84

MLP	0.81	0.70	[0.68 - 0.72]	0.84	0.78	0.68	[0.66 - 0.71]	0.82
-----	------	------	---------------	------	------	------	---------------	------

Notes:

CB: CatBoost

GB: Gradient Boosting

ET: Extra Trees

MLP: Multi-Layer Perceptron

Under the SMOTE-Tomek condition, CatBoost produced a mean macro F1-score of 0.7514, followed by Gradient Boosting at 0.7473 and Extra Trees at 0.7340, while MLP achieved around 0.70. CatBoost also produced the highest mean accuracy at 0.8662. Under SMOTE-ENN, CatBoost and Gradient Boosting produced very similar macro F1 scores, namely 0.7236 and 0.7231. This difference is relatively small compared with each model's confidence interval, so CatBoost's advantage in this strategy is descriptive and should not be interpreted as a significant difference.

Overall, macro F1 under SMOTE-Tomek was higher than under SMOTE-ENN across all classifiers. However, the size of the difference varied across models, indicating that the response to resampling was not entirely uniform. Changes in the number and structure of training samples due to oversampling and cleaning may interact with classifier characteristics, but the internal mechanism causing these differences was not isolated in this study.

To examine changes in classification performance across classes, the CatBoost confusion matrix with SMOTE-Tomek on the held-out test set is shown in Figure 8.

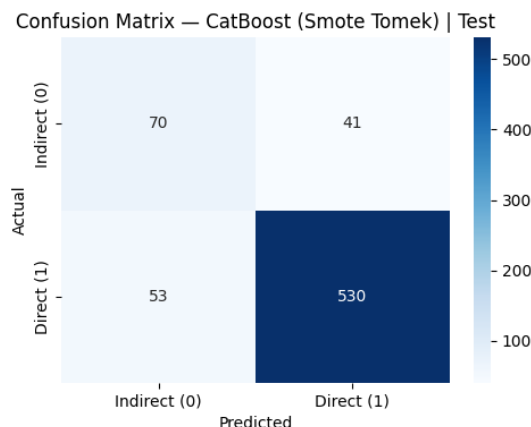


Figure 8. Confusion Matrix CatBoost (Smote Tomek) Test

For the Indirect class, 70 of 111 actual samples were correctly classified (63.1% recall), while 530 of 583 Direct samples were correctly identified (90.9% recall). Compared to the baseline CatBoost model, Indirect recall increased from 0.51 to 0.631, whereas Direct recall decreased from 0.96 to 0.909. These results indicate that SMOTE-Tomek enhances minority-class sensitivity at the expense of majority-class recognition. Thus, resampling operates primarily as a performance trade off between minority sensitivity and

overall model stability rather than a universal enhancement across all metrics. As a diagnostic analysis, Tomek Links and ENN were evaluated independently without SMOTE to contextualize the cleaning component in both hybrid strategies, with results presented in Table VIII.

TABLE VIII
PERFORMANCE OF ENN AND TOMEK LINKS WITHOUT SMOTE

Model	ENN		TOMEK LINKS	
	Accuracy	F1 Macro	Accuracy	F1 Macro
MLP	0.8473	0.7492	0.8905	0.7761
Extra Trees	0.8213	0.7211	0.8631	0.7462
CatBoost	0.8357	0.7364	0.8804	0.7681
Gradient Boosting	0.8646	0.7632	0.8847	0.7756

Tomek Links produced higher accuracy and macro F1-score than ENN across all classifiers. The performance of Tomek Links alone was also relatively close to the baseline in several models. These findings suggest that differences in cleaning mechanisms may contribute to the different responses between SMOTE-Tomek and SMOTE-ENN. However, this analysis is diagnostic, not proof that Tomek Links is universally more effective than ENN or that SMOTE directly causes performance declines. The differences in results may still be influenced by the characteristics of the data distribution and the interaction between resampling and the classifier.

Overall, SMOTE-Tomek showed better aggregate performance than SMOTE-ENN in the experimental setting of this study, but improving sensitivity for the Indirect class did not always lead to better overall generalization. These findings underscore the importance of considering trade-offs across metrics and the characteristics of the resampling strategy, rather than focusing only on the level of class balance in the training set.

D. Model Selection and Hyperparameter Optimization

Based on the 5-fold stratified cross-validation results in Table VII, CatBoost was selected for hyperparameter optimization using macro F1-score as the primary metric. Under SMOTE-Tomek, CatBoost achieved the highest mean macro F1-score of 0.7514, outperforming Gradient Boosting (0.7473), Extra Trees (0.7340), and MLP (0.7048). Under SMOTE-ENN, CatBoost reached 0.7236, marginally surpassing Gradient Boosting (0.7231). Model selection relied strictly on training set cross-validation, keeping the held-out test set isolated to preserve evaluation independence.

Hyperparameter optimization was executed independently for each resampling strategy using Optuna with the Tree-structured Parzen Estimator (TPE) sampler over 100 trials, optimizing for macro F1-score via 5-fold stratified cross-validation (seed 42). Resampling was embedded within the pipeline to process only training folds, preventing synthetic sample generation, data cleaning, or leakage into validation folds. The search space encompassed iterations (200–1500), depth (4–10), log-uniformly sampled learning_rate (0.01–

0.2), l2_leaf_reg (0.1–20), border_count (32–255), random_strength (0–10), and bagging_temperature (0–10), with loss_function set to Logloss and random states fixed at 42 for reproducibility.

TABLE IX
BEST CATBOOST HYPERPARAMETERS OBTAINED FROM OPTUNA

Parameter	SMOTE-Tomek	SMOTE-ENN
iterations	518	1294
depth	4	5
learning_rate	0.0841	0.0383
l2_leaf_reg	0.9619	0.2271
border_count	113	145
random_strength	6.1564	9.9827
bagging_temperature	3.1731	3.4431
Macro F1 (CV)	0.7741	0.7375

The best configuration obtained a cross-validation macro F1-score of 0.7741 with SMOTE-Tomek and 0.7375 with SMOTE-ENN. This difference in configurations shows that optimization produced different CatBoost parameters for each training representation. However, those objective values represent the best performance during the cross-validation process, not a final estimate of the model's generalization ability. The selected configuration was then evaluated on a held-out test set that was not used during either the selection or optimization process.

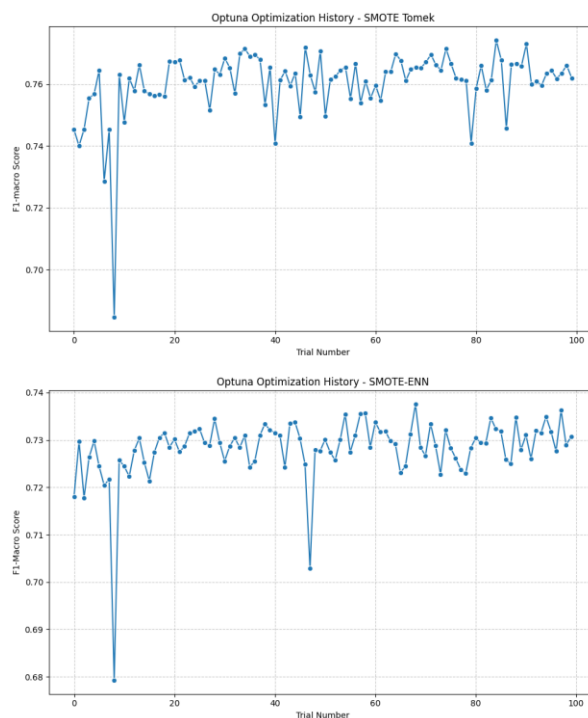


Figure 9. Optuna optimization history for CatBoost

CatBoost evaluation with the optimized configuration on the held-out test set is presented in Table X.

TABLE X
PERFORMANCE OF TUNED CATBOOST ON THE HELD-OUT TEST SET

Resampling strategy	Optuna CV Macro F1	Test Accuracy	Test Macro F1	Test ROC-AUC
SMOTE-Tomek	0.7741	0.8617	0.7549	0.8542
SMOTE-ENN	0.7375	0.8055	0.7164	0.8531

On the held-out test set, CatBoost with SMOTE-Tomek achieved an accuracy of 0.8617, a macro F1-score of 0.7549, and a ROC-AUC of 0.8542. Meanwhile, the SMOTE-ENN configuration achieved an accuracy of 0.8055, a macro F1-score of 0.7164, and a ROC-AUC of 0.8531. Thus, SMOTE-Tomek delivered higher held-out test performance in accuracy and macro F1-score, while the ROC-AUC values of the two strategies were relatively close. This pattern reinforces the earlier finding that SMOTE-Tomek provides higher aggregate performance in the experimental setting of this study.

A comparison between the Optuna objective value and held-out test performance shows a decline in macro F1-score for both strategies. For SMOTE-Tomek, the macro F1-score changed from 0.7741 in cross-validation to 0.7549 on the held-out test set, while for SMOTE-ENN it changed from 0.7375 to 0.7164. This difference shows that the best performance during the optimization process was not fully reflected in the data not used during development. Thus, the Optuna results were used to determine the model configuration, while the held-out test remained the basis for evaluating generalization.

Overall, hyperparameter optimization improved the CatBoost configuration for each resampling strategy, but the results did not change the study's main pattern. SMOTE-Tomek still showed higher held-out test performance than SMOTE-ENN in accuracy and macro F1-score, while the two strategies produced relatively similar ROC-AUC values. These findings indicate that performance variations stem primarily from strategy driven training representations rather than hyperparameter optimization alone.

E. Feature Contribution Analysis

Feature importance analysis was performed on the optimized CatBoost model to evaluate the relative predictive contributions of six physicochemical descriptors under SMOTE-Tomek and SMOTE-ENN strategies, as shown in Figure 10. Under SMOTE-Tomek, A-site electronegativity (ENA) exhibited the highest importance (28.67), followed by B-site electronegativity (ENB: 21.64), A-site ionic radius (RA: 18.88), B-site ionic radius (RB: 13.43), average ionic character (11.50), and electronegativity ratio (ENR: 5.88). Conversely, SMOTE-ENN shifted the top ranking to ENB (20.90), followed by ENA (20.60), average ionic character (17.29), RA (15.90), RB (13.61), and ENR (11.71). Although ENA and ENB remained the dominant descriptors across both setups, shifting training representations altered feature

reliance, noticeably increasing the influence of average ionic character and ENR while reducing that of ENA and RA.

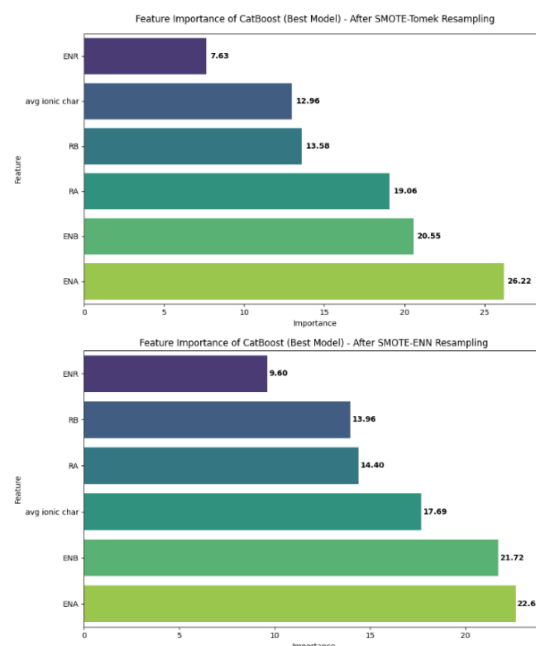


Figure 10. Feature importance of the best CatBoost model

From a materials science perspective, the primary influence of ENA and ENB aligns with their role in governing bonding character, charge redistribution, and electronic structure in ABO_3 materials. However, these importance metrics denote predictive reliance within decision boundaries rather than direct causal drivers, particularly given potential information overlap, such as collinearity between ENA and average ionic character. Overall, these shifts demonstrate that resampling strategies affect not only macro level classification metrics, but also how models dynamically leverage underlying physical descriptors.

IV. CONCLUSION

This study evaluates SMOTE-Tomek and SMOTE-ENN for ABO_3 material band gap classification across MLP, Extra Trees, CatBoost, and Gradient Boosting algorithms, applying resampling exclusively to training data. Baseline models exhibited high overall accuracy (0.8703–0.8934) alongside poor Indirect recall (0.49–0.54) compared to Direct recall (0.94–0.96), underscoring the necessity of macro F1-score and class wise recall metrics. Across all classifiers, SMOTE-Tomek consistently outperformed SMOTE-ENN in macro F1-score. However, resampling introduced notable performance trade-offs; for instance, applying SMOTE-Tomek to CatBoost improved Indirect recall from 0.51 to 0.631 while decreasing Direct recall from 0.96 to 0.909, demonstrating that class balancing optimizes minority-class sensitivity at the expense of overall model stability.

Selected through Optuna hyperparameter optimization based on cross-validation macro F1-scores, CatBoost

combined with SMOTE-Tomek achieved a cross-validation macro F1-score of 0.7741, compared to 0.7375 for SMOTE-ENN. On the held-out test set, SMOTE-Tomek recorded an accuracy of 0.8617, macro F1-score of 0.7549, and ROC-AUC of 0.8542, outperforming SMOTE-ENN, which achieved 0.8055, 0.7164, and 0.8531, respectively. Feature importance analysis identified A-site (ENA) and B-site (ENB) electronegativities as the primary descriptors, with observed shifts in ionic character, RA, RB, and ENR reflecting altered feature reliance under different resampling representations. Overall, combining SMOTE-Tomek with CatBoost provided optimal classification performance for this ABO₃ dataset, proving that effective class balancing requires evaluating cleaning mechanisms and classifier interactions rather than relying solely on equal class counts.

REFERENCES

- [1] H. K. Muhamad Akrom, Supriadi Rustad, Hermawan Kresno Dipojono, Ryo Maezono, "Quantum machine learning for ABO₃ perovskite structure prediction," *Comput. Mater. Sci.*, vol. 250, p. 113694, 2025, doi: 10.1016/j.commatsci.2025.113694.
- [2] S. Touati, A. Benghia, Z. Hebboul, I. K. Lefkaier, M. B. Kanoun, and S. Goumri-Said, "Predictive Machine Learning Approaches for Perovskites Properties Using Their Chemical Formula: Towards the Discovery of Stable Solar Cells Materials," *Neural Comput. Appl.*, vol. 36, no. 26, pp. 16319–16329, 2024, doi: 10.1007/s00521-024-09992-5.
- [3] Z. Chen *et al.*, "Highly Versatile and Accurate Machine Learning Methods for Predicting Perovskite Properties," *J. Mater. Chem. C*, vol. 12, no. 38, pp. 15444–15453, 2024, doi: 10.1039/D4TC02268H.
- [4] B. Wang and J. Wang, "Prediction of Bandgap and Key Feature Analysis of Lead-Free Double Perovskite Oxides Based on Deep Learning," *Molecules*, vol. 31, no. 6, p. 1032, 2026, doi: 10.3390/molecules31061032.
- [5] X.-J. Zhu, N. Zhang, H.-M. Kim, S.-W. Kim, and D.-Y. Jeong, "Data-Driven Design of Tunable Band Gaps in Perovskite Oxides via Machine Learning," *J. Korean Ceram. Soc.*, 2026, doi: 10.1007/s43207-026-00629-y.
- [6] J. Zhang, Y. Li, and X. Zhou, "Machine-Learning Prediction of the Computed Band Gaps of Double Perovskite Materials," 2023. doi: 10.48550/arXiv.2301.03372.
- [7] A. S. Moeini, F. S. Tehrani, and A. Naeimi-Sadigh, "Machine Learning-Enhanced Band Gaps Prediction for Low-Symmetry Double and Layered Perovskites," *Sci. Rep.*, vol. 14, no. 1, p. 26736, 2024, doi: 10.1038/s41598-024-77081-7.
- [8] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.
- [9] B. Nemade, V. Bharadi, S. S. Alegavi, and B. Marakarkandy, "A Comprehensive Review: SMOTE-Based Oversampling Methods for Imbalanced Classification Techniques, Evaluation, and Result Comparisons," *Int. J. Intell. Syst. Appl. Eng.*, vol. 11, no. 9s, pp. 790–803, 2023.
- [10] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data," *ACM SIGKDD Explor. Newsl.*, vol. 6, no. 1, pp. 20–29, 2004, doi: 10.1145/1007730.1007735.
- [11] A. K. Yadav, B. Alam, and O. Pal, "Hybrid Resampling Using SMOTE-ENN and SMOTE-Tomek to Handle Data Imbalance on EEG Signal for Seizure Detection and Classification," *Int. J. Inf. Technol.*, 2025, doi: 10.1007/s41870-025-02832-2.
- [12] G. Husain *et al.*, "SMOTE vs. SMOTEENN: A Study on the Performance of Resampling Algorithms for Addressing Class Imbalance in Regression Models," *Algorithms*, vol. 18, no. 1, p. 37, 2025, doi: 10.3390/a18010037.
- [13] J. Liu *et al.*, "Comparison of Graph Neural Networks and Traditional Machine Learning for Property Prediction in All-Inorganic Perovskite Materials," *Inorganics*, vol. 14, no. 2, p. 58, 2026, doi: 10.3390/inorganics14020058.
- [14] J. O. Ramadhani, A. Z. Rizqi, D. Maharani, and M. Akrom, "Implementation of the Tomek-Link Approach in Machine Learning Models for ABO₃ Perovskite Bandgap Classification," *J. Appl. Informatics Comput.*, vol. 10, no. 2, pp. 1788–1798, 2026, doi: 10.30871/jaic.v10i2.12318.
- [15] G. S. Priyanga, M. N. Mattur, N. Nagappan, S. Rath, and T. Thomas, "Prediction of Nature of Band Gap of Perovskite Oxides (ABO₃) Using a Machine Learning Approach," *J. Mater.*, vol. 8, no. 5, pp. 937–948, 2022, doi: 10.1016/j.jmat.2022.04.006.
- [16] E. F. Swana, W. Doorsamy, and P. Bokoro, "Tomek Link and SMOTE Approaches for Machine Fault Classification with an Imbalanced Dataset," *Sensors*, vol. 22, no. 9, p. 3246, 2022, doi: 10.3390/s22093246.
- [17] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely Randomized Trees," *Mach. Learn.*, vol. 63, no. 1, pp. 3–42, 2006, doi: 10.1007/s10994-006-6226-1.
- [18] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," *Ann. Stat.*, vol. 29, no. 5, pp. 1189–1232, 2001, doi: 10.1214/AOS/1013203451.
- [19] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: Unbiased Boosting with Categorical Features," in *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*, Montréal, Canada: Neural Information Processing Systems Foundation, 2018, pp. 6639–6649.
- [20] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A Next-Generation Hyperparameter Optimization Framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Anchorage, Alaska, United States: Association for Computing Machinery (ACM), 2019, pp. 2623–2631. doi: 10.1145/3292500.3330701.