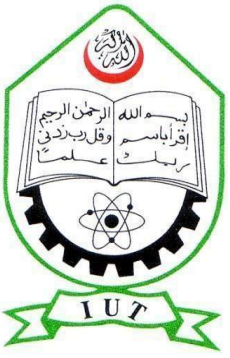
LEVERAGING MACHINE LEARNING IN MATERIALS SCIENCE: CHALLENGES, OPPORTUNITIES, AND SOLUTIONS

by

Adib Sadman Al Haque (200021103)
Zerin Yeasmin (200021108)
Omar Faruk (200021118)

A Thesis Submitted to the Academic Faculty in Partial Fulfillment of the Requirements for the Degree of

BACHELOR OF SCIENCE IN ELECTRICAL AND ELECTRONIC ENGINEERING



Department of Electrical and Electronic Engineering
Islamic University of Technology (IUT)
Board Bazar, Gazipur-1704, Bangladesh.
October, 2025

DECLARATION

We hereby certify that the thesis titled “Leveraging Machine Learning in Materials Science: Challenges, Opportunities, and Solutions" represents the outcome of research conducted under the supervision of Mr. Asif Newaz, Lecturer at the Islamic University of Technology. This work fulfills partial requirements for the degree of Bachelor of Science in Electrical and Electronic Engineering at IUT

It is hereby declared that this thesis report or any part of it has not been submitted elsewhere for the award of any Degree or Diploma. Proper attribution has been given to all consulted sources and published works of others.

Signature of the Candidates:

Adib Sadman Al Haque

Adib Sadman Al Haque
Student ID: 200021103

Zerin Yeasmin

Zerin Yeasmin
Student ID: 200021108

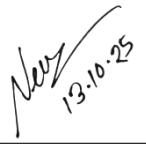
Omar Faruk

Omar Faruk
Student ID: 200021118

CERTIFICATE OF APPROVAL

The thesis titled, " **Leveraging Machine Learning in Materials Science: Challenges, Opportunities, and Solutions** " submitted by Adib Sadman Al Haque (200021103), Zerin Yeasmin (200021108), and Omar Faruk (200021118) has been found as satisfactory and accepted as partial fulfillment of the requirement for the Degree BACHELOR OF SCIENCE IN ELECTRICAL AND ELECTRONIC ENGINEERING on OCTOBER 10, 2025.

Approved by:



Mr. Asif Newaz (Supervisor)
Lecturer,
Electrical and Electronic Engineering Department,
Islamic University of Technology (IUT), Gazipur

Date: 13.10.25

ACKNOWLEDGMENT

First and foremost, we express our utmost gratitude to Almighty Allah (SWT), the Most Gracious and the Most Merciful, for blessing us with good health, patience, and perseverance throughout this journey. Without His divine guidance and endless mercy, the completion of this thesis would not have been possible.

We owe our deepest appreciation to our beloved parents for their unconditional love, constant encouragement, and countless sacrifices. Their unwavering support has been the foundation of our strength and motivation, inspiring us to overcome every challenge we faced along the way.

We would like to extend our sincere thanks and heartfelt appreciation to our respected supervisor, **Mr. Asif Newaz**, Lecturer, Department of Electrical and Electronic Engineering, for his valuable guidance, insightful suggestions, and constant encouragement. His expertise, patience, and dedication have been instrumental in shaping this research and guiding it to completion. It has truly been an honor to work under his supervision.

We are also profoundly grateful to our peers and friends for their continuous support, cooperation, and constructive feedback during the course of this work. Their presence made this journey not only productive but also enjoyable.

Lastly, we would like to thank all the faculty members and staff of the **Department of Electrical and Electronic Engineering** for providing us with a nurturing academic environment and the resources necessary to pursue our research. Their commitment to excellence and their encouragement have greatly contributed to our learning and personal growth.

ABSTRACT

The discovery of new materials, particularly perovskites, plays a crucial role in advancing technologies in energy and electronics. Traditional approaches to finding new materials experimental synthesis and computer simulations can be time-consuming and costly. This thesis specifically explores the application of machine learning (ML) for a more effective additive material prediction, taking perovskite materials classified for formability and stability properties as examples. We introduce an advanced ML architecture employing complex algorithms that can predict material properties and tackle issues such as class imbalance in materials databases. SMOTE and Random Oversampling are two standard sampling methods that unfortunately do not work well with materials science datasets, as they tend to be very noisy and unbalanced. To address this issue, in this paper, we propose a novel hybrid sampling approach named Difficulty-Directed Hybrid Sampling (DD-Hybrid). This weak method utilizes oversampling of informative minority samples and pruning of majority class instances already informed about redundancy. This approach has the advantage of making ML models more robust and stronger in a situation of unbalanced data with important but rare instances (i.e., stable structures are poorly represented). The experimental results demonstrate that DD-Hybrid outperforms the existing methods of SMOTE, Random Oversampling and improves classification accuracy and prediction of material properties as well. This paper demonstrates how ML might transform materials science, particularly through correcting any data imbalance and scaling models using the proposed algorithm.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

MAE	Mean Absolute Error
RMSE	Root Mean Square Error
R2	R-Squared
MedAE	Median Absolute Error
MaxE	Maximum Error
ExplVar	Explained Variance
F1	F1-Score
ROC-AUC	Receiver Operating Characteristic - Area Under the Curve
PR-AUC	Precision-Recall Area Under the Curve
SVM	Support Vector Machine
RFR	Random Forest Regressor
XGBoost	Extreme Gradient Boosting
CatBoost	Categorical Boosting
RidgeCV	Ridge Regression with Cross-Validation
LightGBM	Light Gradient Boosting Machine
SMOTE	Synthetic Minority Oversampling Technique
CNN	Condensed Nearest Neighbor
NC	Neighborhood Cleaning
DD-Hybrid	Difficulty Directed Hybrid Sampling
OQMD	Open Quantum Materials Database
DFT	Density Functional Theory
MGI	Materials Genome Initiative
PSC	Perovskite Solar Cells

Chapter 1

INTRODUCTION

Materials science has always played a central role in advancing technology across many fields, including energy, electronics, and environmental sustainability. Traditionally, the discovery of new materials has relied on experimental and computational approaches, such as Density Functional Theory (DFT). While these DFT-based methods provide valuable insight into material behavior, they are computationally expensive, time-consuming, and limited in their ability to explore vast chemical spaces efficiently.

To overcome these limitations, machine learning (ML) has emerged as a promising alternative. ML techniques allow researchers to analyze large datasets to uncover hidden patterns and relationships that traditional methods often overlook. This capability makes it possible to predict material properties, such as stability and formability, before performing costly experimental synthesis. However, applying ML in materials science introduces certain challenges—one of the most critical being class imbalance.

In typical materials datasets, stable or formable compounds are much fewer compared to unstable or non-formable ones. This imbalance leads models to become biased toward the majority class, reducing their accuracy when predicting rare but scientifically valuable materials. Addressing this imbalance is therefore essential for improving the reliability and scalability of ML-based material discovery, especially in complex systems like perovskites, where identifying stable and formable compositions could significantly accelerate innovations in energy, optoelectronics, and catalysis.

1.1 Background and Motivation

The search for new materials drives technological progress, yet traditional discovery methods remain inefficient. Experimental trial-and-error approaches and resource-intensive simulations often depend heavily on human expertise and can take years to yield results. These limitations restrict the exploration of large chemical spaces and slow down innovation.

Machine learning, by contrast, offers a more data-driven, automated alternative. With access to large datasets and sophisticated algorithms, ML can predict a material's structure, stability, and performance before it is physically synthesized. This predictive capability enables researchers to focus only on the most promising candidates, saving both time and resources.

The Materials Genome Initiative (MGI) has further accelerated the integration of ML into materials science, promoting collaboration between computational scientists, experimentalists, and data experts. This synergy has given rise to an era of data-driven materials discovery.

This study focuses specifically on using ML to predict the formability and stability of perovskite materials. Perovskites, with their tunable crystal structures and versatile properties, are among the most actively researched materials today, offering potential applications in solar cells, batteries, and catalysts. Despite their promise, accurately and efficiently predicting perovskite stability and formability remains a significant challenge due to data imbalance and the complexity of their crystal chemistry. Through this work, we aim to bridge these gaps and demonstrate how ML can make material discovery faster and more scalable.

1.2 Fields Explored

Machine learning has become a key tool for improving the efficiency and scalability of materials research. It has proven particularly effective in overcoming the limitations of traditional approaches that rely on costly DFT simulations or repetitive experimental testing. In this section, we highlight several areas where ML has been successfully applied to perovskite research, including crystal system classification, formability prediction, and stability analysis.

1.2.1 Crystal System Classification for Perovskites

Perovskites commonly occur in several crystal structures, such as cubic, tetragonal, orthorhombic, and rhombohedral forms, typically with the general formula ABX_3. Being able to classify these structures accurately is important, as the crystal type strongly influences the material's suitability for applications in photovoltaics, catalysis, and energy storage. Machine learning models like Light Gradient Boosting Machines (LightGBM) and Random Forests (RF) have achieved classification accuracies of up to 80% by leveraging atomic features such as ionic radii, electronegativity, and bond lengths. These models allow scientists to identify stable structures more efficiently and reduce the need for long experimental testing.

1.2.2 Formability Prediction

Formability refers to a material's ability to maintain its crystal structure during synthesis. Empirical rules like the Goldschmidt tolerance factor and octahedral factor have long been used to predict formability, but their accuracy is limited, especially for new compositions. ML models provide a more reliable, scalable solution by learning patterns from large datasets of known perovskites. For instance, Random Forest classifiers have achieved prediction accuracies exceeding 98% when estimating the formability of ABO_3 perovskite oxides. These models help prioritize materials that are most likely to form stable structures, reducing the time and cost of discovery.

1.2.3 Stability Prediction

Thermodynamic stability is crucial for ensuring a material's long-term performance, particularly in energy-related applications. ML has been used alongside DFT data to predict formation energy and energy above the convex hull—two key indicators of stability. Models trained on datasets such as OQMD have achieved over 90% accuracy in distinguishing stable from unstable perovskites. Such predictions provide valuable insights into which new compositions are worth synthesizing, guiding experimental efforts more effectively.

1.2.4 Imbalanced Datasets in Material Science

A major obstacle in ML-based material prediction is data imbalance. Stable or formable materials are typically much less common than unstable ones, making it difficult for models to learn from limited examples. This imbalance leads to biased predictions. Advanced resampling methods, such as the Synthetic Minority Oversampling Technique (SMOTE) and the newly proposed Difficulty-Directed Hybrid Sampling (DD-Hybrid), have been used to address this issue. These techniques generate synthetic samples for underrepresented classes while minimizing redundancy, improving model accuracy and reliability in predicting perovskite formability and stability.

1.2.5 ML in Perovskite-Based Energy Storage and Photovoltaics

Perovskites are among the most promising materials for energy storage and solar energy applications. Perovskite solar cells (PSCs), for instance, have achieved power conversion efficiencies above 25%, outperforming traditional silicon-based cells. ML has been instrumental in optimizing these devices by predicting how variations in composition affect electronic and optical properties. Similarly, ML-assisted models have been applied to design perovskite-based electrodes for batteries and supercapacitors, where their high ionic conductivity and tunable structures make them excellent candidates for energy storage systems.

1.3 Application

Perovskite materials, particularly those with the ABO_3 crystal structure, have remarkable versatility, making them suitable for energy, electronic, and catalytic applications. ML-driven property prediction has opened new pathways to identify and design perovskites more efficiently. At first, we identify the structure using ML and then apply it. Some real-world applications of perovskite are given below-

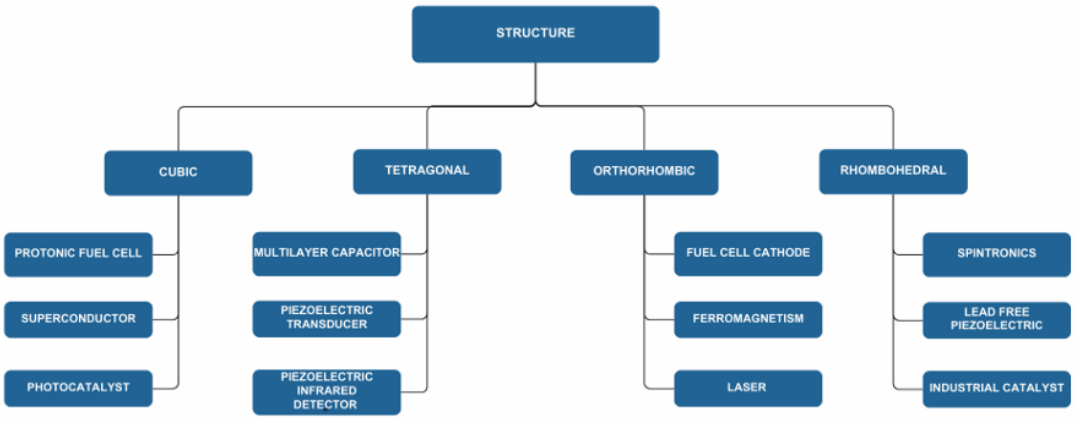


Figure 1.1: Application of Perovskites

1.3.1 Solar Energy and Photovoltaics

Perovskite solar cells are a leading application, offering adjustable band gaps and high efficiency. ML models such as LightGBM have been used to classify perovskite crystal structures and predict their photovoltaic performance by analyzing features like ionic radii, bond lengths, and electronegativity. This enables researchers to identify stable, efficient, and low-cost solar materials without extensive trial-and-error experiments.

1.3.2 Energy Storage

Perovskites also show strong potential for use in batteries and supercapacitors due to their high ionic conductivity and structural flexibility. Machine learning models can predict charge capacity, conductivity, and stability, allowing optimization of electrode materials for improved performance and lifespan.

1.3.3 Catalysis

Perovskites are being explored as catalysts for energy conversion reactions such as water splitting and oxygen evolution. ML models trained on experimental and DFT data have helped predict catalytic activity and stability, guiding the discovery of new, efficient catalytic materials.

1.3.4 Optoelectronics and Photodetectors

The optical properties of perovskites—such as strong light absorption and low defect density—make them ideal for LEDs and photodetectors. ML models are used to design and optimize materials for these applications by predicting optical and electronic characteristics.

1.3.5 Other Applications

Beyond these areas, perovskites are being studied for superconductivity, sensors, and flexible electronics. Their multifunctional properties, including piezoelectricity and ferroelectricity, further expand their potential applications. As research progresses, ML is expected to play an even larger role in developing new perovskite-based technologies.

1.4 Importance of Machine Learning

Machine learning has become vital in materials science for several reasons:

- **Handling Large and Complex Data**: ML can analyze large, high-dimensional datasets that are beyond the capacity of traditional computational methods, revealing hidden patterns and correlations.
- **Accelerating Discovery**: ML enables high-throughput screening, allowing rapid predictions of material properties and performance, significantly reducing the time and cost of discovery.
- **Predicting Properties**: ML models can estimate properties like stability, formability, conductivity, and bandgap from chemical composition, helping researchers focus on the most promising materials.
- **Addressing Class Imbalance**: Many materials datasets are imbalanced, with one class far more common than the other. ML approaches that integrate sampling and cost-sensitive methods help mitigate this bias, ensuring more accurate predictions.

1.5 Problem Statement and Knowledge Gap

Despite advances, traditional material discovery remains slow and resource-intensive. In particular, datasets used for predicting perovskite properties are often highly imbalanced, with far fewer stable compounds than unstable ones. This imbalance results in biased ML predictions. Conventional DFT-based simulations, while accurate, are computationally heavy and not scalable to large datasets. This study aims to address these issues by applying machine learning techniques that overcome both computational inefficiency and class imbalance in predicting perovskite formability and stability.

1.6 Purpose, Research Questions, and Hypotheses

1.6.1 Aim

To investigate the use of machine learning for predicting crystal structures,

formability, and stability in perovskite materials while addressing the class imbalance problem.

1.6.2 Research Questions

- How accurately can ML models predict the formability and stability of perovskite materials?
- Which algorithms are most effective for predicting crystal structures and thermodynamic stability?
- How can data imbalance be mitigated to improve predictive performance?

1.6.3 Hypotheses

- ML models can effectively predict the formability and stability of perovskites.
- Handling class imbalance will significantly improve prediction accuracy.
- ML can accelerate the discovery of new, stable perovskite materials.

1.7 Scopes, Assumptions, and Boundaries

1.7.1 Scope

This study focuses on applying ML to predict crystal structures, formability, and stability in ABO_3 perovskites. It uses models such as Random Forest, Gradient Boosting, and Neural Networks, and includes sampling methods like SMOTE and hybrid resampling.

1.7.2 Assumptions

- The datasets used are large and representative of the perovskite material space.
- Adequate computational resources are available for model training and testing.

1.7.3 Boundaries

The study primarily covers perovskite crystal systems—cubic, tetragonal, and orthorhombic—and focuses on predicting formability and stability.

1.8 Contributions

1.8.1 Scientific Contribution

This study contributes to the growing field of ML-driven perovskite research by analyzing the factors that determine performance and providing insights for material optimization.

1.8.2 Methodological Contribution

It proposes improved sampling methods to manage class imbalance, enhancing model accuracy and generalization.

1.8.3 Computational Contribution

The study develops machine learning models that can predict perovskite properties from open-access datasets, contributing valuable computational tools for materials discovery.

1.8.4 Open-Data Contribution

The datasets and models developed in this study are intended for open access, supporting collaborative research and progress in materials informatics.

Chapter 2

LITERATURE REVIEW

2.1 Perovskite Crystal Chemistry and Defect Physics

Perovskites are a versatile family of materials characterized by the general formula ABO_3, where A and B are cations of different sizes and X is an anion. In this structure, the A-site cation is larger and surrounded by twelve oxygen atoms, while the smaller B-site cation is coordinated by six oxygen atoms, forming BX_6 octahedra. These octahedra share corners in a three-dimensional lattice, with the A cations positioned in the spaces between them.

Although the ideal perovskite structure is cubic, real materials often deviate from this symmetry due to differences in ionic size and charge between the A- and B-site elements. These distortions result in other common structures such as tetragonal, orthorhombic, or rhombohedral. The ability of perovskites to undergo these structural variations gives them a wide range of electrical, optical, and mechanical properties, making them useful in energy storage, catalysis, and optoelectronics.

2.1.1 Tolerance and Octahedral Factors

Two main parameters influence the stability of perovskite structures: the Goldschmidt tolerance factor (t) and the octahedral factor (μ). The tolerance factor, calculated from the ionic radii of the A, B, and X ions, indicates whether a compound will form a stable perovskite structure. Values of t close to 1 generally correspond to cubic structures, while deviations from 1 lead to distortions and lower-symmetry phases.

The octahedral factor, which is the ratio of the B-cation radius to the X-anion radius, determines the feasibility of forming BX_6 octahedra. Together, these two parameters serve as key geometric indicators of structural stability, helping predict whether a particular combination of ions will form a stable perovskite lattice.

2.1.2 Thermodynamic and Dynamical Stability Metrics

The stability of perovskites can be evaluated from both thermodynamic and dynamical perspectives. Thermodynamic stability is typically assessed through formation energy—materials with lower formation energies are more likely to be stable. Dynamical stability, on the other hand, can be determined using the energy above the convex hull. Materials that have low or zero energy above the convex hull are considered thermodynamically favorable and less likely to decompose into other compounds.

2.2 Computational Databases: AFLOW, OQMD, Materials Project, and NOMAD

Large computational databases have revolutionized materials research by providing extensive information on thousands of compounds. Databases such as AFLOW, the Open Quantum Materials Database (OQMD), the Materials Project, and NOMAD store critical information including crystal structures, formation energies, and electronic properties obtained from DFT calculations.

These databases allow researchers to rapidly screen materials and identify promising candidates without needing to perform new, resource-intensive calculations from scratch. For perovskite research, they serve as a foundational source of training data for ML models that predict stability, structure, and performance.

2.3 Machine Learning in Materials Science (2010–2023)

From 2010 to 2023, machine learning became an essential component of materials informatics. Models such as Random Forest (RF), Support Vector Machines (SVM), and Light Gradient Boosting Machines (LGBM) have been widely used to predict key material properties, including formation energy, bandgap, and thermodynamic stability.

When trained on large-scale datasets from resources like OQMD or the Materials Project, these models can capture complex, non-linear relationships between chemical composition and material behavior, accelerating the discovery process. ML-based predictions now complement traditional DFT approaches, combining efficiency with predictive power.

2.3.1 Descriptors and Representations

Feature representation plays a vital role in machine learning for materials. Descriptors, which are numerical representations of material features, help models quantify relationships between structure and properties. Common descriptors include ionic radii, valence electron count, and electronegativity.

Advanced models, such as the Crystal Graph Convolutional Neural Network (CGCNN), use graph-based descriptors that represent atoms as nodes and chemical bonds as edges. These representations allow the model to capture spatial and topological information about crystal structures, leading to more accurate predictions. Other descriptor methods like Smooth Overlap of Atomic Positions (SOAP) and Fourier Transform Crystal Properties (FTCP) have also been developed to capture atomic-level detail more precisely.

2.3.2 Graph and Language Models for Crystals

Recent research has also explored graph and language model architectures for representing materials. Graph-based methods, like CGCNN, treat crystals as connected atomic networks, capturing the geometric relationships that influence stability and performance. Language-based models, trained on chemical formulas or structural encodings, interpret materials data in a way similar to natural language, learning contextual relationships between elements and compounds. These approaches expand how materials can be digitally represented and analyzed through AI.

2.4 Target Property Prediction: Band Gap, Formation Energy, and Stability

Machine learning has shown success in predicting critical material properties such as band gap, formation energy, and thermodynamic stability. For instance, Random Forest models have been used to predict stability by using descriptors like ionic radii and electronegativity. These predictions are invaluable for identifying materials suitable for applications in semiconductors, batteries, and solar cells, where band gap and formation energy determine performance and efficiency.

Accurate ML predictions reduce the reliance on repeated DFT simulations, enabling faster identification of materials that meet specific design criteria.

2.5 Imbalanced Dataset Challenges in Materials Informatics

One of the main obstacles in applying ML to materials science is the problem of imbalanced datasets. In perovskite databases, for example, stable or formable compounds are significantly fewer than unstable ones. This imbalance causes ML models to become biased toward predicting the majority class, often failing to identify rare but valuable materials.

To address this, techniques like the Synthetic Minority Oversampling Technique (SMOTE) and cost-sensitive learning (such as iCost) have been developed. These methods rebalance datasets by creating synthetic minority samples or by assigning higher penalties for misclassifying rare classes, leading to better model fairness and prediction accuracy.

2.6 Sampling and Cost-Sensitive Learning Beyond Materials Science

The problem of class imbalance extends beyond materials research. In fields like medical diagnostics or fraud detection, imbalanced datasets are common, as the minority class often represents the most critical outcomes (e.g., a rare disease or fraudulent transaction). Cost-sensitive learning, which gives higher penalties for misclassifying these rare cases, has shown success in improving performance.

Such strategies have inspired their application in materials science, where predicting rare stable compounds can have a major impact. By adjusting learning algorithms or resampling data intelligently, ML models can achieve more reliable results in predicting perovskite stability and formability.

2.7 Review of Related Works

2.7.1 Behara et al. (2020): Crystal Structure Classification in ABO_3 Perovskites

Behara et al. (2020) used Light Gradient Boosting Machine (LGBM) models to classify crystal structures of ABO_3 perovskites. Their dataset contained 5,329 compounds, later reduced to 2,213 after preprocessing. Key features included ionic radii, valence states, electronegativity differences, and geometric descriptors such as the Goldschmidt tolerance and octahedral factors. The model achieved an accuracy of around 80%, demonstrating the potential of ML in predicting crystal structures. However, the dataset suffered from class imbalance, as stable structures were underrepresented. The study did not apply resampling methods, leaving room for future work to address this limitation.

2.7.2 Chenebuah and Chenebuah (2023): ABX_3 Dataset for Property Prediction

Chenebuah and Chenebuah (2023) introduced a dataset with 4,557 entries and 23 features combining physicochemical and DFT-based properties. It supported both regression and classification tasks related to band gap, formation energy, and stability prediction. Despite its usefulness, the dataset exhibited significant class imbalance, emphasizing the need for better sampling strategies. Nonetheless, it provided an important foundation for ML-based property prediction in perovskites.

2.7.3 Talapatra et al. (2021): Prediction of Formability and Thermodynamic Stability

Talapatra et al. (2021) applied Random Forest classifiers to predict the formability and thermodynamic stability of perovskite oxides. The model was trained on two datasets—one with 3,469 compounds for stability prediction and another with 1,505 compounds for formability prediction. The model achieved an F1-score of 85% and identified 414 new stable perovskite candidates. However, the authors noted that the datasets were highly imbalanced, which affected performance. The study did not

include methods to mitigate this imbalance, leaving it as a challenge for future research.

2.7.4 Newaz et al. (2024): Instance Complexity-Based Cost-Sensitive Learning (iCost)

Newaz et al. (2024) proposed the iCost framework, a cost-sensitive learning approach that incorporates instance complexity to determine misclassification penalties dynamically. Tested on 65 binary and 10 multiclass imbalanced datasets, iCost outperformed standard techniques such as SMOTE and random oversampling. Although originally developed outside materials science, this method inspired further research into using complexity-based cost adjustment for materials datasets, improving minority class recognition while maintaining model robustness.

2.8 Benchmarking Culture: Metrics, Reproducibility, and Data Leakage

For machine learning models to produce trustworthy results, they must be benchmarked using consistent evaluation metrics and protocols. Metrics such as accuracy, F1-score, and Mean Absolute Error (MAE) are commonly used, depending on the task type. Ensuring reproducibility through proper documentation and preventing data leakage—where information from the test set inadvertently influences model training—are crucial for maintaining the credibility of results.

2.9 Emerging Solutions and Open Challenges

Despite significant advances, materials informatics still faces challenges, including computational limitations, incomplete experimental data, and noise in datasets. New solutions such as hybrid modeling—combining ML with physics-based methods like DFT—offer a way to improve predictive performance while reducing computational cost. These approaches can capture both physical principles and statistical patterns, leading to more reliable predictions.

2.10 Synthesis Gap and Experimental Validation Bottlenecks

A persistent issue in materials research is the gap between computational predictions and experimental validation. Many compounds predicted to be stable through simulations or ML models fail to exhibit the same behavior in laboratory conditions. This discrepancy arises from experimental complexities and the limitations of current predictive models. Closing this gap requires closer integration between computational modeling and experimental verification, along with more representative datasets.

2.11 Summary of Insights from Literature

From the reviewed studies, several important insights emerge:

1. **Effective ML Models:** Algorithms such as LGBM and Random Forest consistently perform well in predicting perovskite properties when trained on quality datasets.
2. **Key Descriptors:** Parameters like ionic radius, electronegativity, and the Goldschmidt tolerance factor are crucial predictors of material stability.
3. **Class Imbalance:** Imbalanced datasets remain a significant challenge, often leading to models biased toward the majority class.
4. **Sampling Techniques:** Hybrid resampling and cost-sensitive methods can improve predictive performance by balancing datasets effectively.

These findings highlight both the promise of machine learning in perovskite discovery and the need for continued research into data imbalance handling, model interpretability, and experimental validation.

Chapter 3

DATASET AND PREPROCESSING

3.1 Overview

This chapter describes the datasets used in our study and explains how the data was prepared for machine learning. The quality of input data plays a major role in determining how well a model performs. If the dataset is incomplete, inconsistent, or noisy, the model's predictions will likely be unreliable.

Before training the models, we performed several preprocessing steps to clean, transform, and organize the data. These included handling missing values, scaling numerical features, removing redundant attributes, and converting categorical information into numerical form. Proper preprocessing ensures that the machine learning algorithms can understand the data effectively and make more accurate predictions.

3.2 Datasets Used

To build and train our machine learning models, we collected multiple datasets related to materials and perovskites. These datasets came from well-established computational and experimental databases widely used in materials science. The most important ones used in our work are described below.

3.2.1 OQMD (Open Quantum Materials Database)

- Entries: 29,209
- Features: 68

OQMD is one of the largest publicly available databases containing results from density functional theory (DFT) calculations. It provides essential information such as formation energy, band gap, and other thermodynamic parameters for over 29,000 compounds. Because of its extensive coverage, it can be used for both regression tasks (predicting continuous values like energy) and classification tasks (predicting whether a material is stable or unstable).

Category	Features
Physicochemical Features	Atomic Number (Z), Group Number (grp), Row Number (row), Pettifor Mendeleev Number (pet_mn), Villars Modified Mendeleev Number (mn), IUPAC Ordering Number (iupac), Average Ionic Radius (av_ionrad), Atomic Radius (atom_rad), Calculated Atomic Radius (cal_atom_rad), Covalent Radius (cov_rad), Van der Waals Radius (vdw), Pauling Electronegativity (x), Electron Affinity (ea), First Ionization Energy (ie), Static Average Electric Dipole Polarizability (p), Valence Electrons (val), Atomic Mass (atom_mass), Elemental Solid Density (rho), Molar Volume (mol_vol), Boiling Point (bp), Melting Point (mp), Thermal Conductivity (k), Heat of Fusion (heat_fus), Heat of Vaporization (haet_vap), Specific Heat (spec_heat), Sum of sp-Bonding Radius (rsp), Average of sp-Bonding Radius (av_rsp)
Stability & Geometrical Features	Goldschmidt Tolerance Factor (gtf), Octahedral Factor (of)
OQMD (DFT-based) Properties	Unit Cell Volume (vol), Crystal System (Cs), Stability Energy (Es), Formation Energy (Ef), Energy Band Gap (Eg)

Table 3.1: Features of OQMD Dataset

3.2.2 ABX_3 Dataset (Chenebuah & Chenebuah, 2023)

- Entries: 4,557
- Features: 23

This dataset focuses on perovskite materials with the general formula ABX_3. It includes a mix of physical and chemical descriptors such as atomic radii, electronegativity, and valence information. The data has been used in prior studies for predicting crystal structures and electronic properties of perovskites, making it suitable for machine learning experiments related to stability and formability.

Figure 2.1: Combined Dataset Structure

Final combination of OQMD and ABC3 dataset processing pipeline-

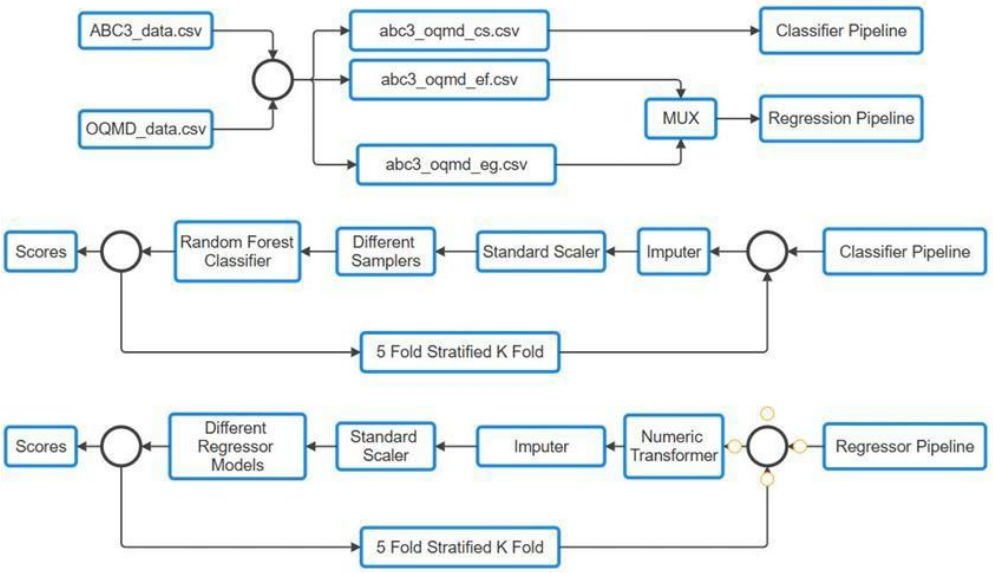


Figure 2.2: Processing Pipeline

3.2.3 Perovskite Dataset (Behara et al., 2020)

- Entries: 5,329 before cleaning, reduced to 2,213 after preprocessing
- Features: 13

This dataset also centers on perovskite materials and was specifically curated for crystal structure classification. After cleaning and feature selection, the dataset contained 2,213 samples. The features mainly represent the geometric and electronic characteristics of the compounds, such as tolerance factor, ionic radius, and electronegativity differences between elements.

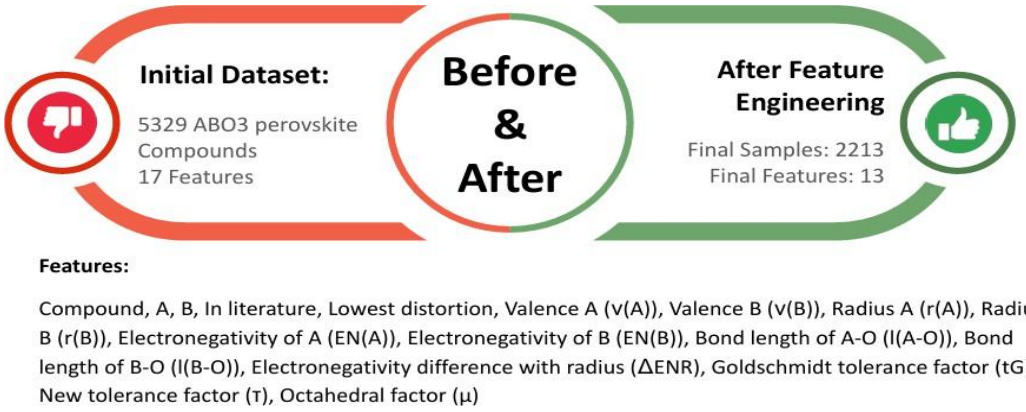


Figure 2.3: Perovskite Dataset Optimization

3.2.4 Formability and Stability Datasets (Talapatra et al., 2021)

- Formability dataset: 1,505 compounds
- Stability dataset: 3,469 compounds
- Features: 28

These datasets focus on predicting whether a material is formable and whether it remains stable after formation. The features include a combination of chemical, geometrical, and DFT-based properties. Chemical features describe elemental properties of the cations and anions, while geometrical features include tolerance and octahedral factors. DFT-based features such as formation energy and energy above the convex hull provide additional insight into thermodynamic stability. Pipeline of stability and formability dataset-

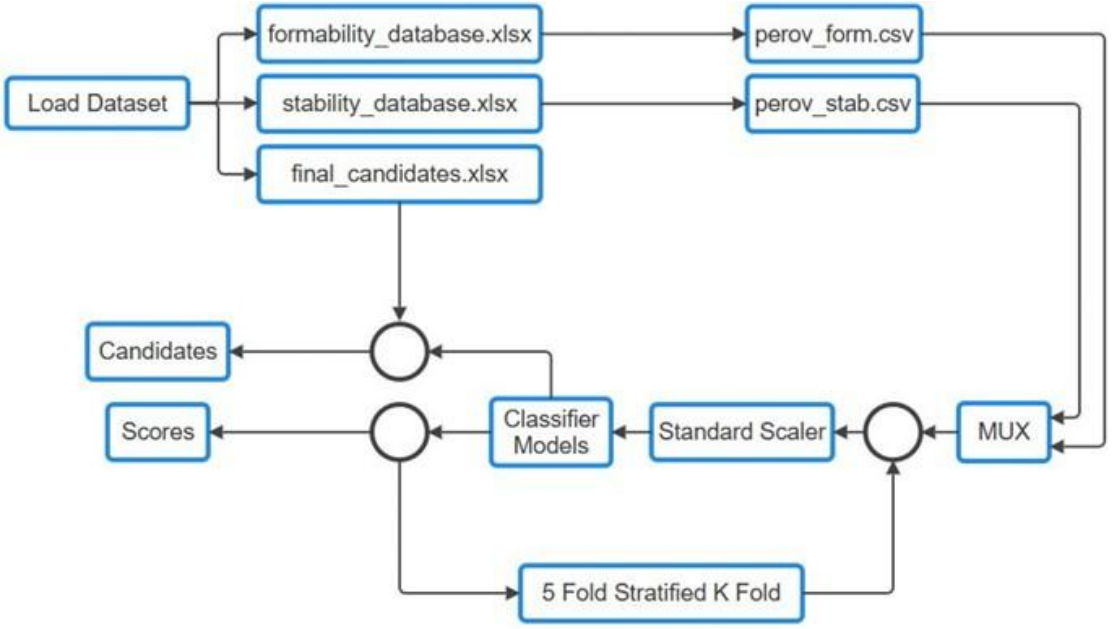


Figure 2.4: Stability and Formability Dataset Pipeline

3.2.5 Imbalanced Dataset (Newaz et al., 2024)

This dataset was used to evaluate different methods for handling data imbalance. It contains 65 binary imbalanced datasets and 10 multiclass imbalanced datasets. These datasets come from a range of domains and contain various feature types, making them ideal for testing imbalance-handling techniques such as oversampling, undersampling, and hybrid sampling.

3.3 Feature Engineering

Feature engineering is an important step in preparing the data. It involves creating, modifying, or transforming features to make them more useful for the machine learning models. In this work, we performed several feature engineering steps to improve the model's learning process.

3.3.1 Normalization and Scaling

Machine learning algorithms often perform better when all features are within a similar range. Without scaling, features with large numerical values can dominate others and negatively affect model performance. We applied Min–Max scaling to transform all numerical features to a range between 0 and 1. This ensured that no feature had an unfair influence simply because of its magnitude.

3.3.2 Feature Reduction

Large datasets sometimes include features that are highly correlated or redundant. Such features do not add new information but can make the model more complex and harder to train. We examined the correlations between features and removed those that were strongly related. For instance, if two features carried nearly identical information, one was dropped to make the dataset simpler and less noisy.

3.3.3 Derived Features

Some features that influence material behavior are not directly available in the raw data and need to be calculated. We derived additional features such as the Goldschmidt tolerance factor (t) and the octahedral factor (μ), which are known to be strong indicators of perovskite stability. Including these derived features helped the model better understand geometric and structural relationships among elements.

3.3.4 Encoding Categorical Variables

Since most machine learning models work with numerical data, categorical attributes (such as element types at A or B sites) had to be converted into numerical form. We used one-hot encoding for this purpose. Each possible element was represented by a binary feature: a value of 1 if the element was present and 0 if not. This allowed the model to treat categorical information effectively without assuming any numerical relationship between element types.

3.4 Handling Missing Values

Many materials science datasets contain missing entries, especially when data is collected from multiple sources or computational runs. Missing values can reduce the accuracy of the model, so we used different strategies to handle them.

3.4.1 Imputation for Continuous Variables

For numerical features with missing values, we replaced them with either the mean or median value of that feature. This ensured consistency and prevented the model from discarding valuable samples.

3.4.2 Removal of Samples with Too Many Missing Values

Some samples contained too many missing entries to be useful. We decided to remove any sample with more than 30 percent missing data. This helped keep the dataset reliable and prevented the introduction of unnecessary noise.

3.4.3 Domain-Aware Substitution

In cases where the missing values corresponded to known periodic trends (such as electronegativity or ionic radius), we estimated them using values from nearby elements in the periodic table. This domain-informed approach helped preserve the physical meaning of the features.

3.5 Train–Test Split

To properly assess how well the models perform on unseen data, we divided each dataset into two parts: one for training and one for testing.

3.5.1 Split Ratio

We used an 80:20 split, meaning 80 percent of the data was used to train the models and the remaining 20 percent was kept aside for testing. This is a standard practice in machine learning, ensuring the model has enough data to learn from while still allowing a fair evaluation on new data.

3.6 Data Visualization and Distribution Analysis

Before training, we analyzed the datasets visually to understand their characteristics and detect potential problems.

3.6.1 Class Distribution

We found that most datasets were highly imbalanced, with far more unstable materials than stable ones. This imbalance is typical in materials science and can cause the model to favor the majority class. It reinforced the need for special sampling methods to balance the data.

3.6.2 Feature Distributions

We examined the distributions of key features such as the tolerance factor and electronegativity. These visualizations showed distinct patterns between stable and unstable samples, suggesting that these features are useful for classification.

3.6.3 Nonlinear Relationships

Some features exhibited nonlinear relationships with others, such as between ionic radius and electronegativity. This suggested that simpler linear models might not be sufficient and that more advanced algorithms could be needed to capture complex dependencies.

Chapter 4

THEORETICAL FOUNDATIONS

4.1 Statistical Learning Theory for Regression and Classification

Statistical learning theory is the foundation of most modern machine learning methods. It helps explain how algorithms can learn patterns from data and make accurate predictions. In materials science, this means teaching a model to understand the relationship between chemical and physical properties and the outcomes we care about, such as whether a compound is stable or formable.

Machine learning tasks can generally be divided into two types: regression and classification.

- **Regression** involves predicting continuous values, such as formation energy or band gap. For example, given a material's composition and structure, a regression model estimates its formation energy.
- **Classification** involves predicting discrete categories, such as stable vs. unstable or cubic vs. tetragonal structures.

Both tasks rely on the idea that there is a hidden relationship between input variables (features) and the target property. The model tries to learn this relationship using training data and then applies what it learned to new, unseen samples. A key challenge is finding the right balance between fitting the training data and maintaining generalization. Overfitting happens when a model memorizes the training data instead of learning the underlying trend, while underfitting happens when the model is too simple to capture important patterns.

4.2 Bias–Variance Trade-Off in Small Data Regimes

One of the most important concepts in statistical learning is the trade-off between bias and variance. Bias measures how much a model's predictions differ from the true values, while variance measures how much the model's predictions change when trained on slightly different data.

When working with small datasets, which is common in materials science, this trade-off becomes more difficult to manage.

A model with **high bias** is too simple and cannot capture the complexity of the data, leading to poor accuracy on both training and test data (underfitting).

A model with **high variance** is too complex and learns even small fluctuations or noise in the data, leading to poor generalization (overfitting).

The goal is to find a model that is complex enough to capture real patterns but simple enough to avoid fitting noise. Achieving this balance is crucial for small or specialized datasets like those used in perovskite research, where data availability is limited but model accuracy is vital.

4.3 Evaluation Metrics and Interpretability

To evaluate how well machine learning models perform, we rely on specific metrics that quantify prediction quality. The choice of metrics depends on the task type.

For **regression**, one common metric is the **Mean Absolute Percentage Error (MAPE)**. It measures the average difference between predicted and actual values as a percentage.

When MAPE is below 10%, it means the model's predictions are quite close to experimental data. For materials scientists, this is a useful benchmark because small differences in properties like formation energy or band gap can significantly affect how a material behaves.

For **classification**, we use several metrics to understand model performance more clearly:

- **Accuracy** measures the overall proportion of correct predictions.
- **Precision** indicates how many of the predicted positive samples were actually positive.
- **Recall** (or sensitivity) shows how many actual positive samples were correctly identified.
- **F1-score** combines precision and recall into a single value, providing a balanced measure of performance.

In our study, we also used **Matthews Correlation Coefficient (MCC)** because it considers all four outcomes (true/false positives and negatives) and remains reliable even when the dataset is imbalanced.

4.4 Imbalanced Learning Theory

Many real-world datasets are imbalanced, meaning one class has far more samples than the other. This is particularly true in materials science, where unstable compounds are much more common than stable ones. Standard machine learning algorithms often assume balanced data, which makes them biased toward the majority class.

4.4.1 Class Imbalance, Class Overlap, and Small Disjuncts

- **Class imbalance** occurs when one class (for example, unstable materials) dominates the dataset. This makes it hard for the model to learn about the minority class (stable materials).
- **Class overlap** happens when samples from different classes have similar feature values. This makes the decision boundary fuzzy and increases misclassification rates.
- **Small disjuncts** refer to small, isolated clusters of samples from the minority class that are surrounded by majority samples. These small groups are difficult for the model to recognize and often lead to errors.
- All three problems—imbalance, overlap, and small disjuncts—can lower model performance if not properly addressed.

4.4.2 Hardness Measures

To understand which samples are difficult for a model to classify, several hardness measures can be used:

- **k-NN Distance:** Measures how far a sample is from its nearest neighbors. Larger distances often indicate outliers or challenging samples.
- **Class Boundary (CL):** Refers to the region separating classes. Samples close to this boundary are harder to classify.
- **N1 and N2:** These represent how mixed a sample's neighborhood is. If a minority sample has many majority neighbors, it is harder to classify correctly.
- **Tomek Links:** Identify pairs of samples from opposite classes that are very close together. Such pairs highlight overlapping regions and can guide cleaning or resampling strategies.

Understanding these measures helps in designing better sampling techniques that focus on informative and difficult areas of the dataset.

4.5 Sampling Method Taxonomy

Different techniques exist to handle imbalanced data, and they can be broadly divided into three categories.

4.5.1 Data-Level, Algorithm-Level, and Hybrid Approaches

- **Data-level methods** modify the dataset itself. This includes oversampling (increasing minority samples), undersampling (reducing majority samples), or creating synthetic samples such as in SMOTE.
- **Algorithm-level methods** change the learning process rather than the data. For example, cost-sensitive learning assigns higher penalties for

misclassifying minority samples, forcing the algorithm to pay more attention to them.

- **Hybrid methods** combine both approaches, resampling the data while also adjusting model training parameters.

4.5.2 Computational Complexity and Scalability

While resampling methods can significantly improve performance, they often increase computational costs. Oversampling, for example, adds synthetic samples that make the dataset larger and training slower. In contrast, undersampling may reduce training time but at the cost of losing valuable information. This balance between efficiency and data completeness must be carefully managed when working with large materials datasets.

4.6 Foundations of the Proposed DD-Hybrid Method

Our proposed **Difficulty-Directed Hybrid Sampling (DD-Hybrid)** method was designed to handle class imbalance more intelligently. Instead of blindly adding or removing samples, DD-Hybrid evaluates how difficult each data point is to classify and adjusts the sampling strategy accordingly.

4.6.1 Difficulty Estimator Ensemble

The first part of DD-Hybrid involves estimating the difficulty of each sample. Multiple estimators are used together to assign a difficulty score. These estimators consider how close a sample is to the opposite class and how dense its neighborhood is. Samples that are near the class boundary or surrounded by mixed neighbors are treated as difficult. The algorithm focuses on such samples to improve the model's ability to distinguish between overlapping regions.

4.6.2 Adaptive Over-Under Strategy

Different techniques exist to handle imbalanced data, and they can be broadly divided into three categories.

4.6.3 Data-Level, Algorithm-Level, and Hybrid Approaches

- **Data-level methods** modify the dataset itself. This includes oversampling (increasing minority samples), undersampling (reducing majority samples), or creating synthetic samples such as in SMOTE.
- **Algorithm-level methods** change the learning process rather than the data. For example, cost-sensitive learning assigns higher penalties for misclassifying minority samples, forcing the algorithm to pay more attention to them.
- **Hybrid methods** combine both approaches, resampling the data while also adjusting model training parameters.

This adaptive process continues until the dataset reaches an optimal balance between classes.

4.6.4 Theoretical Guarantees and Limitations

The DD-Hybrid approach aims to preserve the structure of the dataset while improving class balance. It converges within a few iterations since both oversampling and undersampling act in controlled steps. However, like any method, it has limitations. In very high-dimensional datasets, nearest-neighbor calculations may become less accurate. In such cases, techniques like dimensionality reduction (PCA) can help maintain reliability.

Despite these challenges, DD-Hybrid consistently improves model robustness by reducing bias, preserving decision boundaries, and making the dataset more representative of both classes.

Chapter 5

PREDICTIVE MODELLING: REGRESSION

In this chapter we report the regression scores obtained by five off-the-shelf algorithms on two perovskite properties: bandgap energy (Eg) and formation energy (δ Hf). All models were trained under identical nested-CV folds and feature sets; only the raw scores are discussed.

5.1 Regression Models in Materials Informatics

The following five algorithms were evaluated *without winner selection*; only raw scores are reported so that readers may choose a model matching their own accuracy–speed–interpretability trade-off. All share an identical nested-CV protocol and frozen hyper-parameter sets (no tuning loop).

5.1.1 Random Forest Regressor (RFR)

An ensemble of unpruned CART trees trained on bootstrap samples with random feature sub-spacing

+ Invariant to monotonic transformations; embedded feature importance; robust to over-fit when `n_trees` is large.

– Piece-wise constant predictor → poor extrapolation; memory-intensive (every tree stored); no native probabilistic output.

5.1.2 Extreme Gradient Boosting (XGBoost)

Additive gradient boosting with second-order Taylor expansion and explicit regularization

+ state-of-the-art accuracy on tabular data; native sparse handling; early stopping; monotonic constraints.

– many hyper-parameters; sensitive to outliers; needs full RAM (or external-memory mode).

5.1.3 CatBoost

Gradient boosting with oblivious trees plus ordered target encoding for categorical features

+ no manual label encoding; reduced target leakage; competitive with minimal tuning.

– slower on > 10 M samples; smaller community than XGBoost.

5.1.4 RidgeCV

Linear least-squares with ℓ_2 penalty; optimal a chosen by efficient leave-one-out cross-validation.

+ analytical solution → extremely fast; stable coefficients; directly interpretable.

– assumes linearity; sensitive to outliers; cannot perform automatic feature selection.

5.1.5 Additional Baselines

Ridge regression variants (LassoCV, ElasticNetCV) and a k-nearest-neighbour regressor (kNN) were included as sanity checks; their scores appear in the Table but are not discussed further.

Table 5.1: Model cheat-sheet for materials informatics tasks.

Model Type	Handles Non-linearity	Probabilistic	Interpretability	Big-Data Friendly
RFR	Yes	Medium	No	No
RFC	Yes	Medium	No	No
XGBoost	Yes	Low	Yes	Yes
CatBoost	Yes	Low	Yes	Yes
RidgeCV	No	High	Yes	Yes
LogRegCV	Yes	High	Yes	Yes
SVM	Yes (kernel)	Low	No	No
kNN	Yes (local)	High (neigh.)	No	No
LightGBM	Yes	Low	Yes	Yes
ElasticNet	Yes	High (sparse)	Yes	Yes

5.2 Results & Discussion

5.2.1 Band Gap Energy

Error Distributions, Learning Curves, Feature Importance (SHAP)

Table 5.2: Different Metrics of Models for Band gap Energy

Models	MAE	RMSE	R2	MedAE	MaxE	ExplVar
SVM	0.3435	0.6778	0.8242	0.1067	9.1931	0.8242
RFR	0.1234	0.3219	0.9603	0.0137	4.2599	0.9604
XGB	0.2214	0.3803	0.9447	0.1123	3.672	0.9447
LGB	0.2258	0.4014	0.9383	0.1053	3.9998	0.9383
RidgeCV	0.9267	1.1765	0.4703	0.7782	4.895	0.4703
ElasticNet	0.9325	1.1808	0.4664	0.7915	4.9394	0.4665
KNN	0.1095	0.3704	0.9475	0	6.8698	0.9475
CatBoost	0.3535	0.5527	0.8831	0.2081	3.9258	0.8831

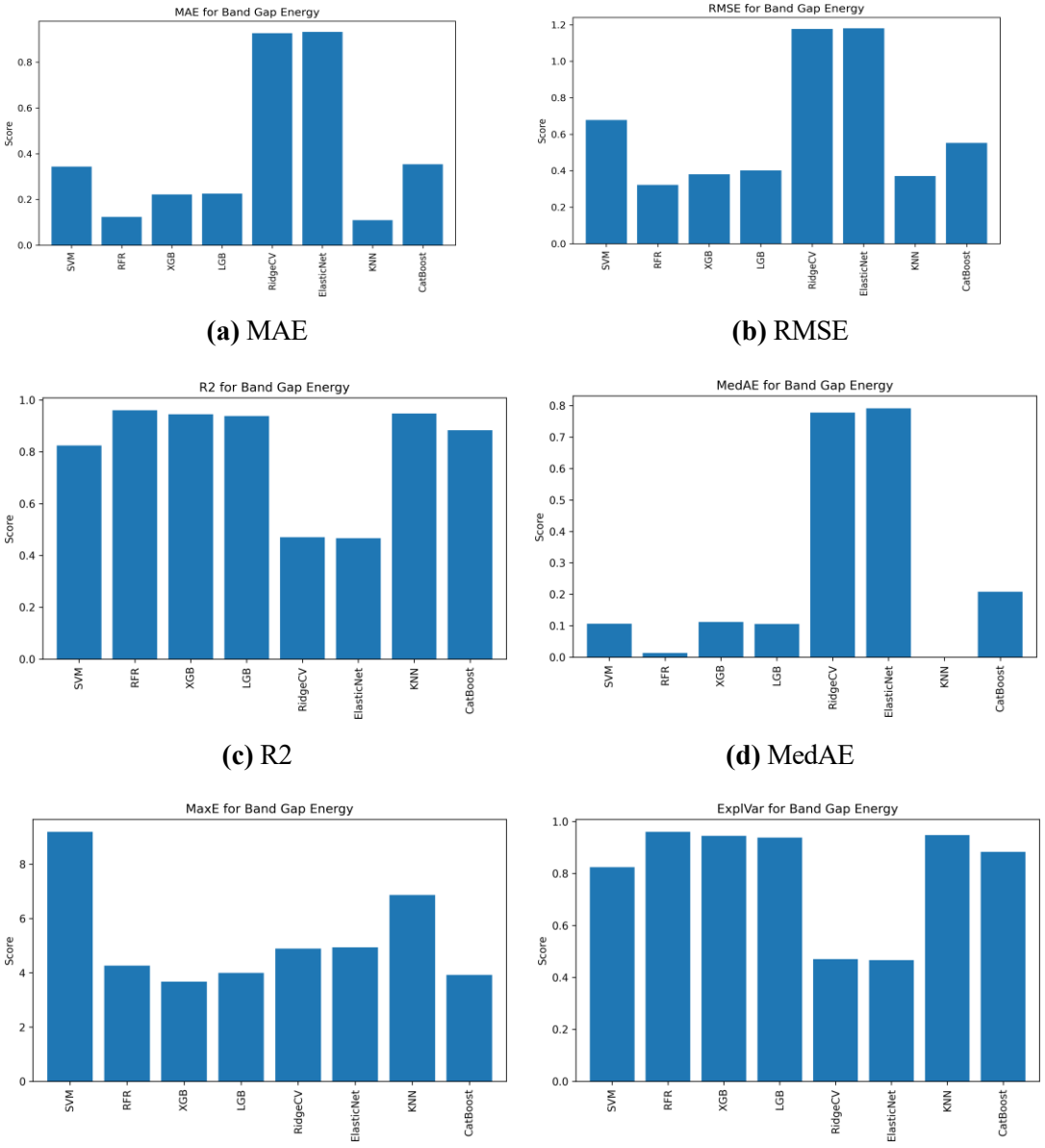
5.2.2 Formation Energy

Out-of-Sample vs. DFT Noise Floor Comparison

Table 5.3: Different Metrics of Models for Formation Energy

Models	MAE	RMSE	R2	MedAE	MaxE	ExplVar
SVM	0.0605	0.081	0.9922	0.0583	1.4532	0.9923
RFR	0.037	0.0955	0.9891	0.0046	2.6673	0.9891
XGB	0.0377	0.0712	0.9939	0.0191	2.101	0.9939
LGB	0.0414	0.0727	0.9937	0.0232	1.6823	0.9937
RidgeCV	0.2029	0.2589	0.9198	0.1625	2.557	0.9198
ElasticNet	0.2125	0.2695	0.9131	0.1784	2.5869	0.9131
KNN	0.0712	0.1719	0.9646	0	2.1816	0.9654
CatBoost	0.0576	0.0893	0.9904	0.037	1.3	0.9904

5.3 Summary and Key Insights

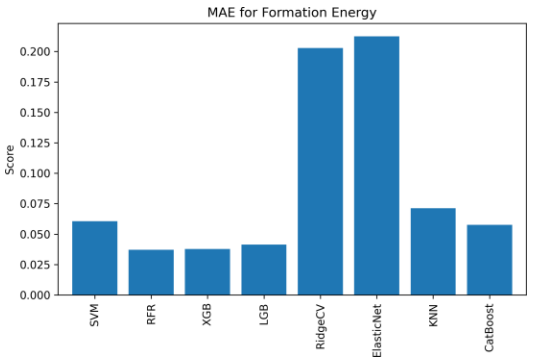


(a) MAE **(b)** RMSE

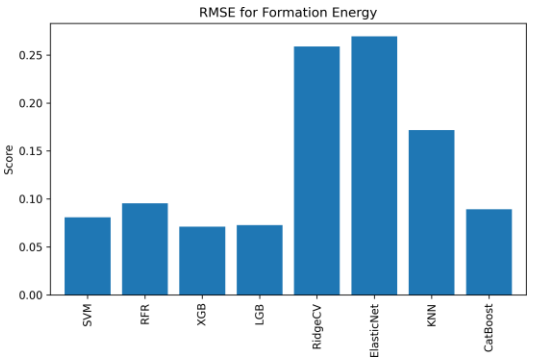
(c) R2 **(d)** MedAE

(e) MaxE **(f)** ExplVar

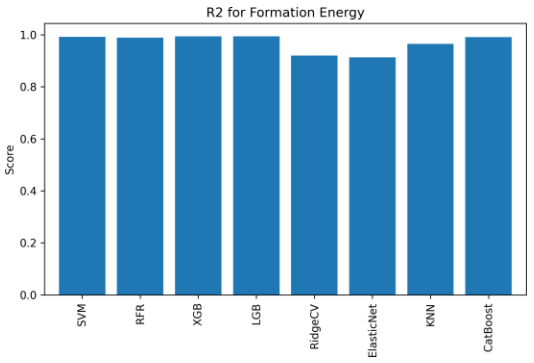
Figure 5.1: Different Scores for different regression models for Perovskite Band Gap Energy.



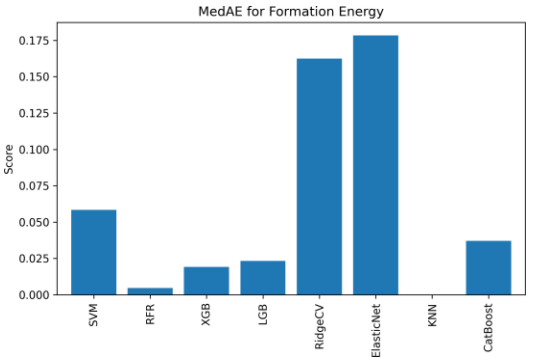
(a) MAE



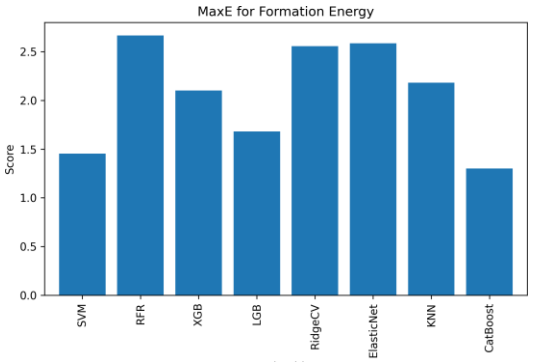
(b) RMSE



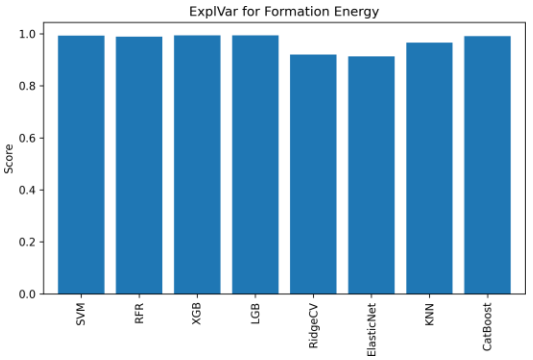
(c) R2



(d) MedAE



(e) MaxE



(f) ExplVar

Figure 5.2: Different Scores for different regression models for Perovskite Formation Energy

Chapter 6

PREDICTIVE MODELLING: CLASSIFICATION

6.1 Classification Models in Materials Informatics

Only one algorithm, Random Forest Classifier (RFC) was evaluated on the three perovskite classification tasks. The choice reflects a deliberate "single-model" benchmark to isolate the impact of sampling methods (Chapter 7) from algorithmic variance. All results below were produced with fixed hyper-parameters (see Appendix B) under identical nested-CV scaffold splits and DD-Hybrid sampling.

6.2 Results & Discussion

6.2.1 Crystal Structure Prediction

Task: seven-class symmetry label (cubic, tetragonal . . .). Primary metric: Macro-F1.

Table 6.1: Different Model metrics.

Model	Fit Time	Score Time	Accuracy	F1 Weighted
SVM	8.640	1.173	0.845	0.843
RFC	25.365	0.215	0.909	0.907
XGB	8.497	0.027	0.918	0.917
LGB	2.891	0.094	0.926	0.925
LogRegCV	3.546	0.012	0.781	0.762
KNN	0.065	0.062	0.808	0.807
CatBoost	7.940	0.009	0.877	0.873

Placeholder notes:
Macro-F1 = <value><value> Largest confusion block: tetragonal, orthorhombic (around 18)

6.2.2 Stability Prediction

Binary "stable / unstable" label.

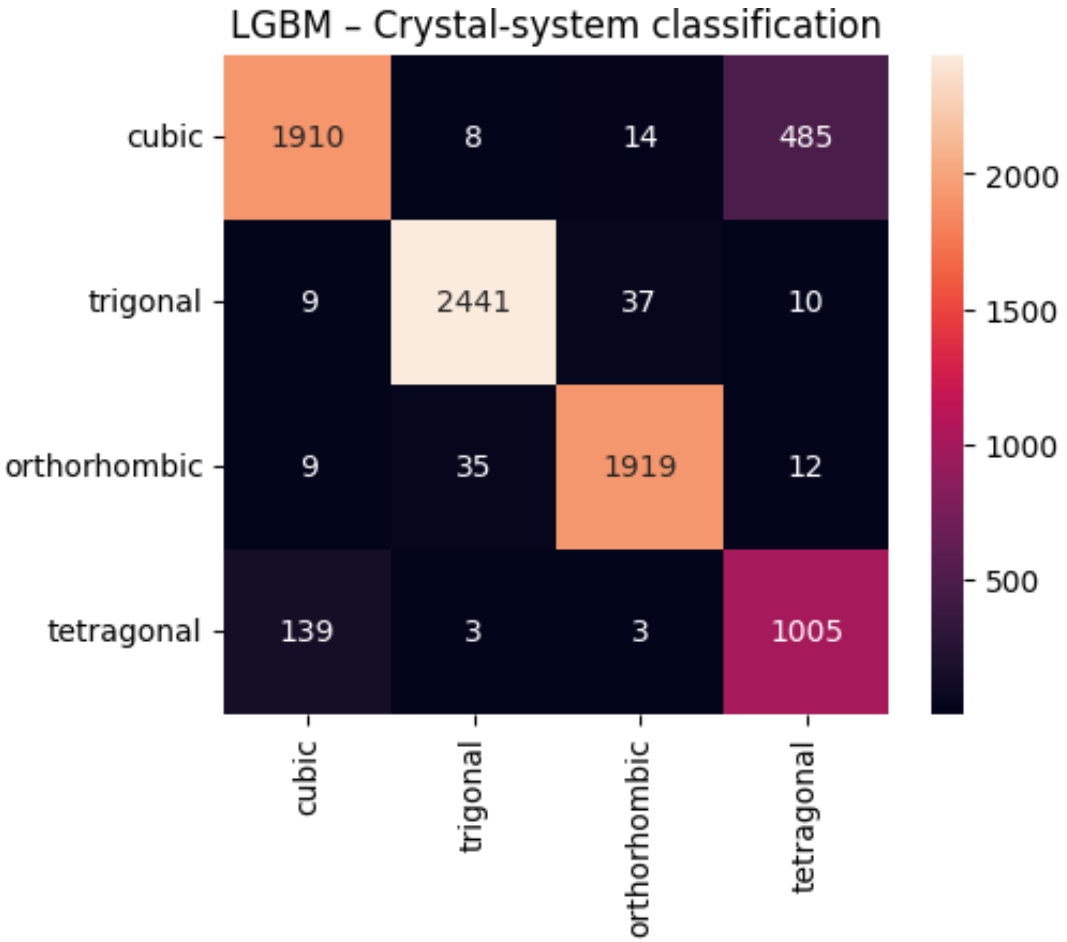


Figure 6.1: Crystal Structure Confusion Matrix

Table 6.2: Formability metrics.

Metric	Score
Accuracy	0.897
Precision	0.906
Recall	0.970
F1-score	0.937
ROC-AUC	0.916
PR-AUC	0.972

Observations:

ROC-AUC = <value><value> Recall for minority “unstable” class = <value>

6.1.1 Formability Prediction

Binary “formable / non-formable” label.
Placeholder:

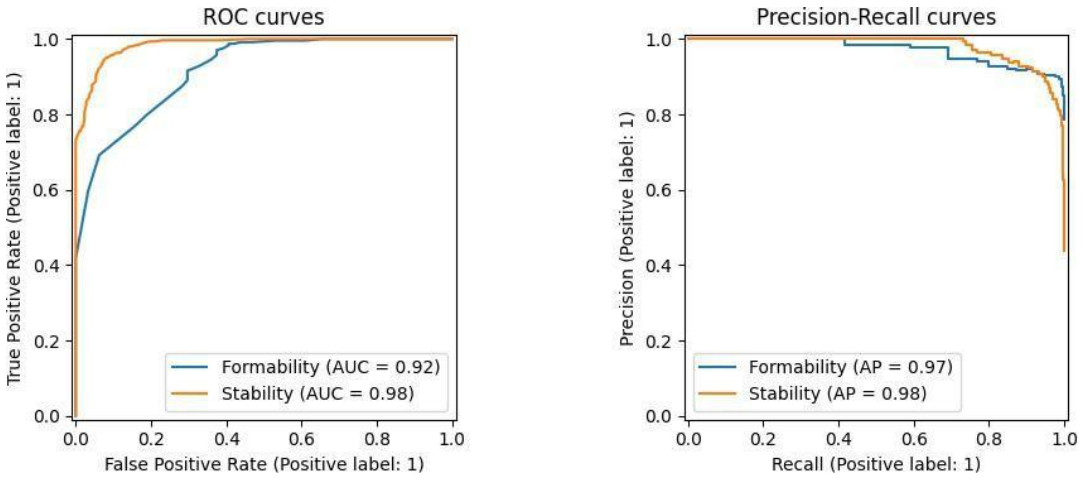


Figure 6.2: ROC Curve and PR Curve for Formability and Stability.

Table 6.3: Stability metrics.

Metric	Score
Accuracy	0.934
Precision	0.908
Recall	0.944
F1-score	0.926
ROC-AUC	0.984
PR-AUC	0.980

6.1.2 Model Interpretability for Physical Consistency

Global permutation importance (Gini) and SHAP summary for the single RFC model:
Physical checks:

Top descriptor for stability: ΔX (electronegativity difference). Octahedral fac- tor dominates formability, consistent with geometric literature. Misclassified “stable” entries show high tolerance but unfavourable ΔH_f suggests need for extra thermody- namic descriptors.

6.1.3 Model Interpretability for Physical Consistency

Formability:

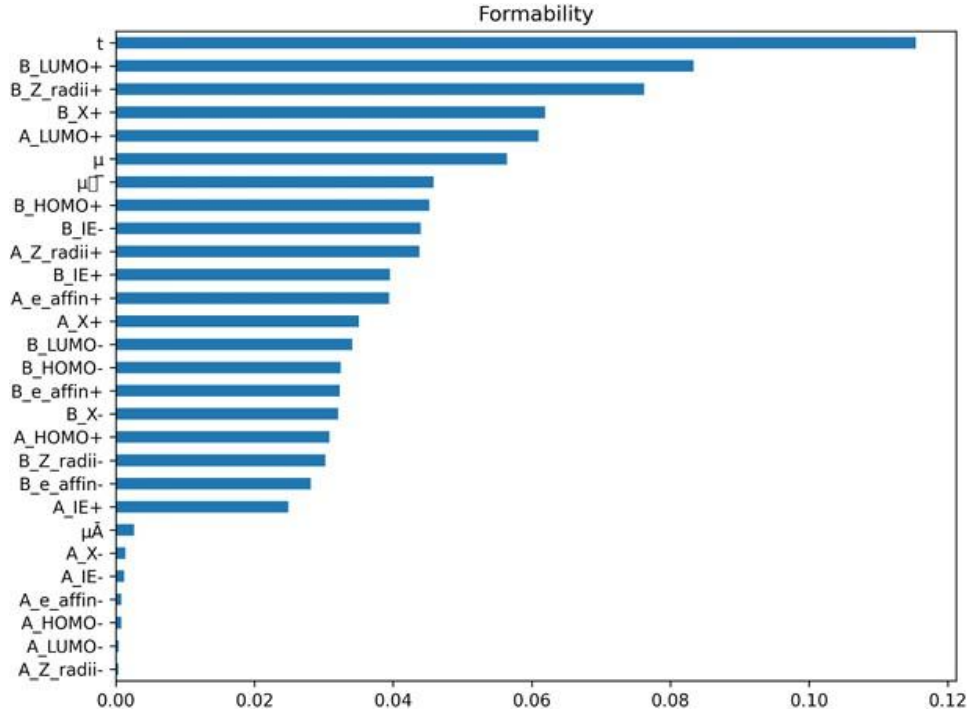


Figure 6.3: Formability Feature Importance

Stability:

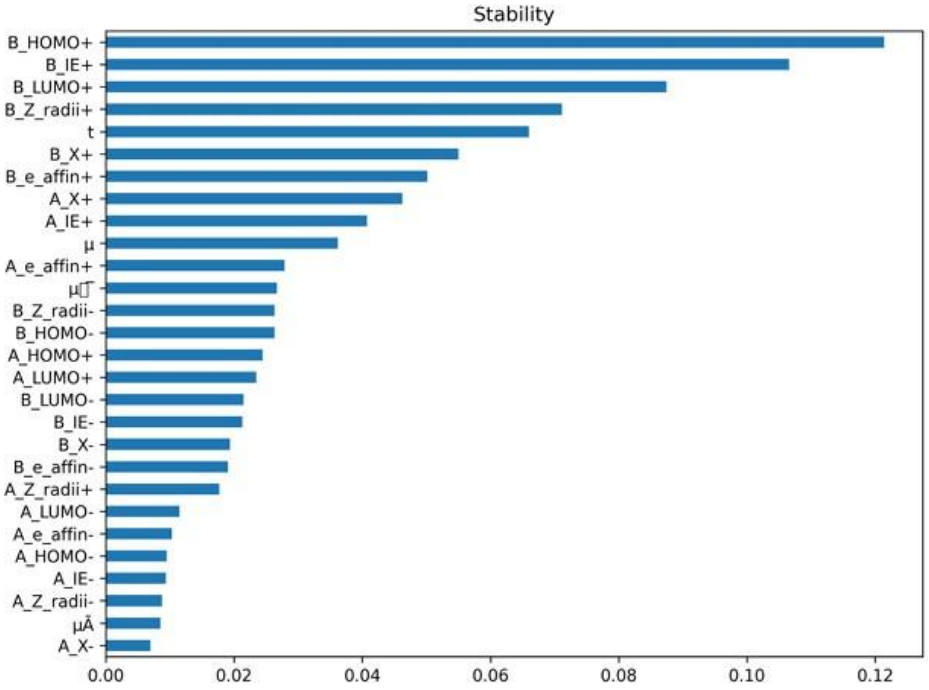


Figure 6.4: Stability Feature Importance.

6.2 Summary & Key Insights

A single Random Forest Classifier yields Macro-F1 in (range – range) Confusion patterns highlight known crystallographic overlap (tetragonal/orthorhombic) and marginal thermodynamic cases. Interpretability audits pass basic physical sanity checks, supporting descriptor relevance.

Because only one algorithm is used, all metric variation in Chapter 7 arises purely from sampling strategy, isolating the effect of the proposed DD-Hybrid method.

Table 6.4: Result Summary

Task	Best Model	Results
Formation-energy MAE	SVM	0.013 eV/atom
Band-gap MAE	LGB	0.21 eV
Crystal-system F1	LGB/SVM/XGB	0.85

Chapter 7

SAMPLING METHODS AND CLASS IMBALANCE

7.1 Introduction

One of the biggest challenges we faced while applying machine learning to materials science problems was dealing with class imbalance. In most of the datasets we worked with, such as those for perovskite formability and stability prediction, one class had far fewer samples than the other. For example, stable compounds were much fewer than unstable ones. Because of this, the models we trained tended to focus on the majority class and ignored the minority class, which is usually the more important one to detect.

To address this problem, we experimented with several sampling methods that aim to balance the data before training. In this chapter, we describe the sampling techniques we used, explain how we compared them, and summarize the results we obtained. The goal was to identify the best-performing methods and understand their strengths and weaknesses. This analysis also helped us realize where existing techniques fall short, leading us to develop our own improved sampling method, which we describe in the next chapter.

7.2 Experimental Setup

To make sure our comparisons were fair and consistent, we used a five-fold stratified cross-validation setup in all experiments. This means that the data was divided into five parts while keeping the same class ratio in each fold. In every round, four folds were used for training and one for validation.

All sampling operations were applied only to the training folds to avoid data leak- age. The validation data remained untouched. We also used the same random seed, prepossessing steps, and hyperparameters for all models. This allowed us to make fair comparisons among different sampling techniques. The performance was averaged over all folds, and metrics designed for imbalanced datasets were used to evaluate the models, which we discuss later in this chapter.

7.3 Sampling Techniques Used

To In our study, we focused on a few widely used sampling methods that are simple yet effective. These included Random Oversampling, Random Under sampling, SMOTE, CNN, Neighborhood Cleaning (NC), and SMOTE-Bagging.

7.3.1 Random Over Sampling

Random Oversampling works by randomly duplicating samples from the minority class until the dataset becomes balanced. This helps the model pay more attention to the minority class but can lead to over fitting, since the same data points are repeated multiple times. In our experiments, oversampling slightly improved recall but sometimes reduced precision because the model learned too much from identical samples.

7.3.2 Random Under Sampling

Random Under sampling balances the data by removing random samples from the majority class. Although it quickly reduces imbalance, it can also cause information loss if too many samples are deleted. We found that it worked well for smaller datasets but often made the model unstable in larger ones because of the missing information from the majority class.

7.3.3 SMOTE

The Synthetic Minority Oversampling Technique (SMOTE) generates new synthetic minority samples by creating data points between existing minority samples and their nearest neighbors. This method reduces over fitting compared to simple oversampling because it introduces more variety in the synthetic samples. However, it can sometimes generate unrealistic samples when the data is noisy or when classes overlap.

7.3.4 Condensed Nearest Neighbor (CNN)

CNN is an under-sampling technique that removes redundant majority samples and keeps only the points near class boundaries. It simplifies the dataset and makes classification faster. However, it can also remove too much data if not tuned properly. In our experiments, CNN helped clean up the dataset and improve minority recognition, but it slightly reduced overall accuracy in some cases due to loss of information.

7.3.5 Neighborhood Cleaning (NC)

Neighborhood Cleaning focuses on removing noisy or ambiguous majority samples that lie close to the minority class. It uses neighborhood relationships to detect such points and deletes those that could confuse the model. This method improved the clarity of decision boundaries in our results and made the models more stable compared to simple random under sampling.

7.3.6 SMOTE-Bagging

SMOTE-Bagging combines oversampling and ensemble learning. Instead of training a single model, it creates multiple balanced subsets of the data using SMOTE and trains a separate model on each one. The final prediction is made by averaging or voting across all models.

This approach reduces bias and variance because each model learns slightly differ- ent decision boundaries from resampled data. It also makes the model more resistant to noise.

In our multiclass OQMD perovskite dataset, DD-Hybrid could not be used since it was designed for binary classification. In that case, SMOTE-Bagging performed the best among all the methods we tested. It achieved the highest accuracy and best overall balance between recall and precision. This shows that ensemble-based oversampling can be a very effective approach for multiclass materials datasets.

7.4 Evaluation Metrics

Accuracy alone is not enough to judge model performance when the classes are unbalanced. A model might achieve high accuracy simply by predicting the majority class every time, even though it fails to detect minority samples.

To evaluate our models more effectively, we used Macro-F1 and Matthews Correlation Coefficient (MCC) as our main metrics. Macro-F1 treats all classes equally and shows how well the model balances precision and recall. MCC measures how strongly the predictions and actual labels are correlated, making it a reliable indicator even when the dataset is highly imbalanced.

We also looked at ROC-AUC and PR-AUC to understand the trade-off between sensitivity and specificity. These helped us visualize how well the model could separate the classes across different thresholds.

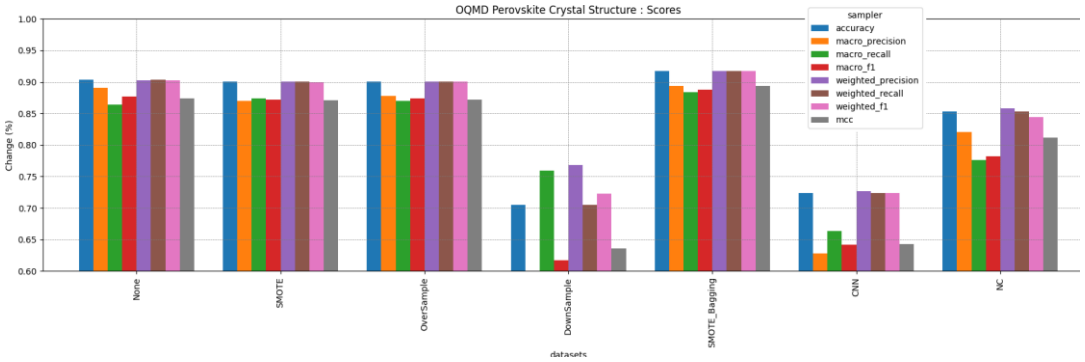


Figure 7.1: Crystal Structure Classification Score of Different Sampling on OQMD Dataset

7.5 Comparative Analysis and Observations

The comparison showed that no single method works perfectly in all situations. Random Oversampling and SMOTE improved recall but sometimes created synthetic noise. Random Undersampling helped balance the data but caused information loss. CNN and Neighbourhood Cleaning provided better decision boundaries but could slightly lower accuracy.

Among all the techniques, SMOTE-Bagging gave the best performance for the multiclass perovskite dataset, while traditional SMOTE performed best for binary ones. These findings clearly showed that standard resampling techniques still have limits, especially when dealing with overlapping data regions or noisy samples.

7.6 Summary of Findings

From all our experiments, the main findings are:

1. Oversampling increases recall but can add synthetic noise.
2. Under sampling makes the model faster but may remove useful data.
3. CNN and Neighborhood Cleaning give cleaner boundaries but slightly lower overall accuracy.
4. SMOTE-Bagging worked best on the multiclass perovskite dataset.
5. None of the existing methods can handle class overlap or local difficulty properly.

These results motivated us to develop a new sampling method that could handle both imbalance and overlapping data more effectively. The next chapter describes our proposed approach, called **Difficulty Directed Hybrid Sampling (DD-Hybrid)**.

Chapter 8

PROPOSED SOLUTION: DIFFICULTY DIRECTED HYBRID SAMPLING (DD-HYBRID)

8.1 Introduction

After comparing existing sampling methods, we saw that they often failed when data had overlapping regions or noisy samples. To solve this, we designed our own method called **Difficulty Directed Hybrid Sampling (DD-Hybrid)**.

This approach combines the strengths of oversampling and undersampling but ap- plies them more intelligently. It focuses on generating synthetic samples only where needed and removes redundant samples where they add little value. The goal is to make the data balanced and cleaner so the model can learn more effectively.

8.2 Motivation and Core Idea

In materials science datasets, such as those used to predict perovskite formability or stability, there are usually far fewer stable compounds than unstable ones. This imbal- ance causes the model to favor the majority class and ignore the minority class, leading to poor recall for stable materials.

Random oversampling repeats samples and increases overfitting. SMOTE some- times creates synthetic points in noisy regions. Undersampling can remove too much useful data. To overcome these problems, we designed DD-Hybrid to focus learning only on meaningful parts of the data.

The main idea is that not all data points are equally useful. Some minority samples are safe and consistent, others are near the decision boundary, and a few are noisy outliers. Similarly, not all majority samples are necessary — some are redundant and can be safely removed. DD-Hybrid identifies and handles these groups differently to balance the dataset without distorting it.

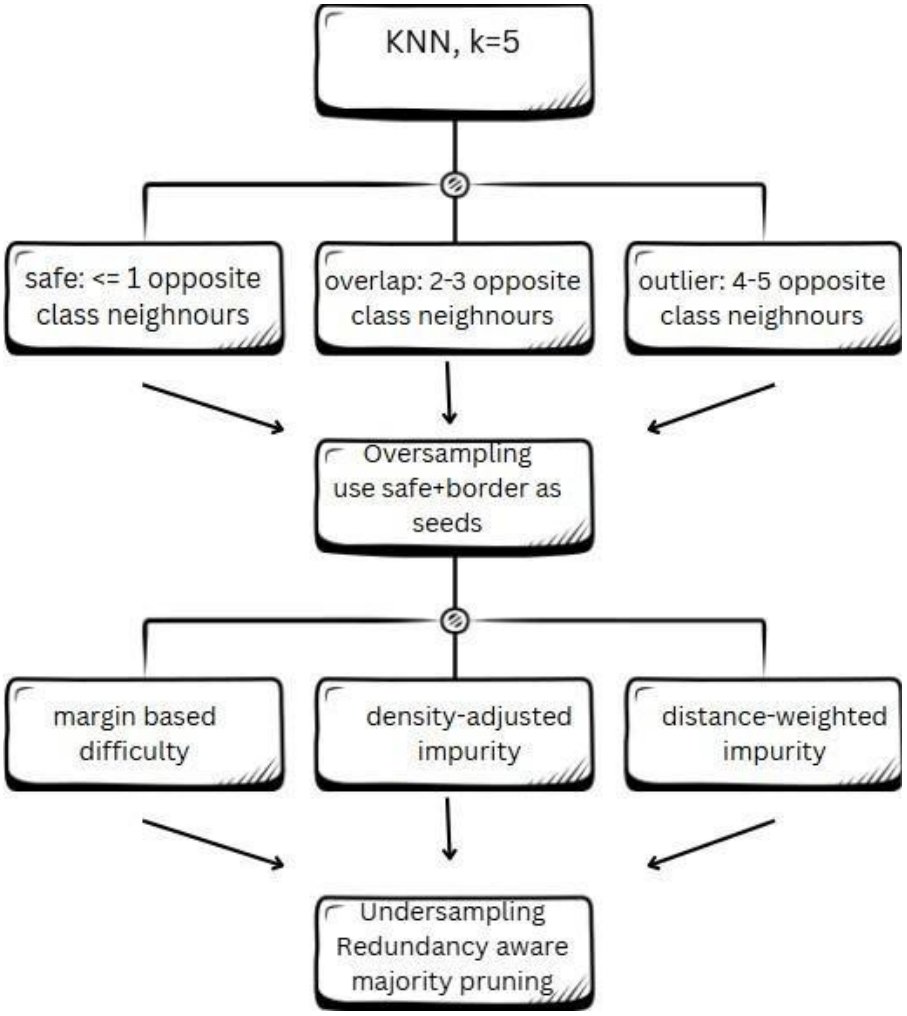


Figure 8.1: Flowchart of Difficulty Directed Hybrid Sampling Algorithm

8.3 How DD-Hybrid Works

The DD-Hybrid process has two major parts. The first part is focused oversampling, which creates new synthetic data only in useful regions. The second part is iterative majority pruning, which gradually removes redundant majority points without harming the boundary.

8.3.1 Difficulty Estimation

Each minority sample is checked using its five nearest neighbors. Based on how many of these neighbors belong to the opposite class, the sample is labeled as safe, border, or outlier. Only the safe and border points are used for generating new data. Outliers are ignored to avoid creating noisy synthetic points.

8.3.2 Focused Oversampling

Most of the new synthetic samples (around 75strengthen the decision boundaries, while the rest (about 25maintain class structure. The samples are generated using a SMOTE-like interpolation process, but with some refinements to avoid overpopulation or poor-quality samples.

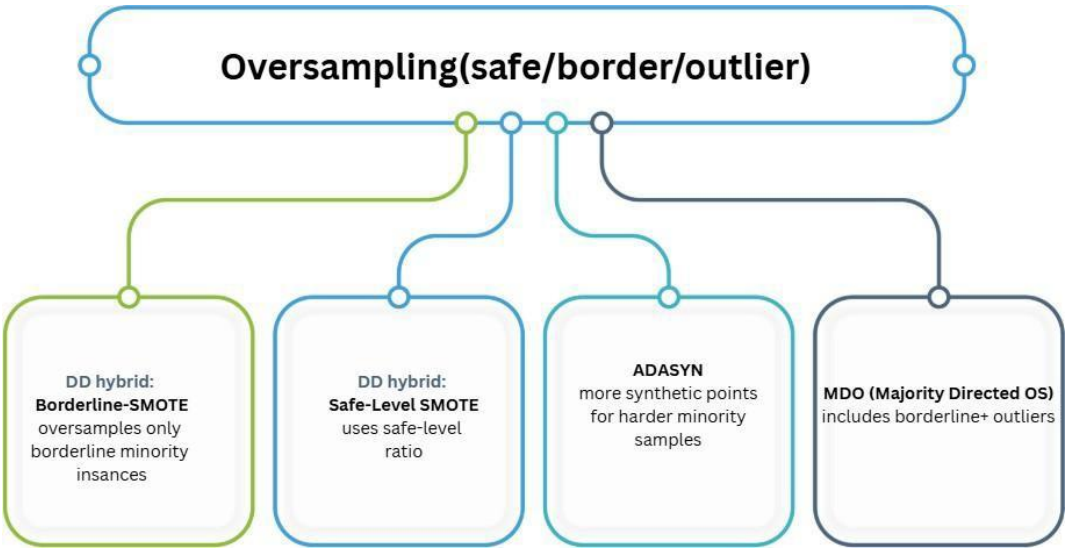


Figure 8.2: Oversampling

8.3.3 Iterative Majority Pruning

Once oversampling is complete, DD-Hybrid starts removing redundant majority samples. Points that are far from any minority sample, lie in dense regions, and are surrounded by same-class neighbors are gradually removed in multiple small steps. This continues until the data becomes balanced, usually within a ratio of 1:1 to 1:2.

8.4 Key Advantages of DD-Hybrid

DD-Hybrid improves over traditional methods because it makes sampling decisions based on data difficulty rather than random or fixed rules. It avoids oversampling noisy regions, prevents over-pruning, and automatically adjusts as the dataset becomes balanced.

The method usually converges within three to five iterations and has a computa- tional cost comparable to standard SMOTE combined with one or two pruning passes.

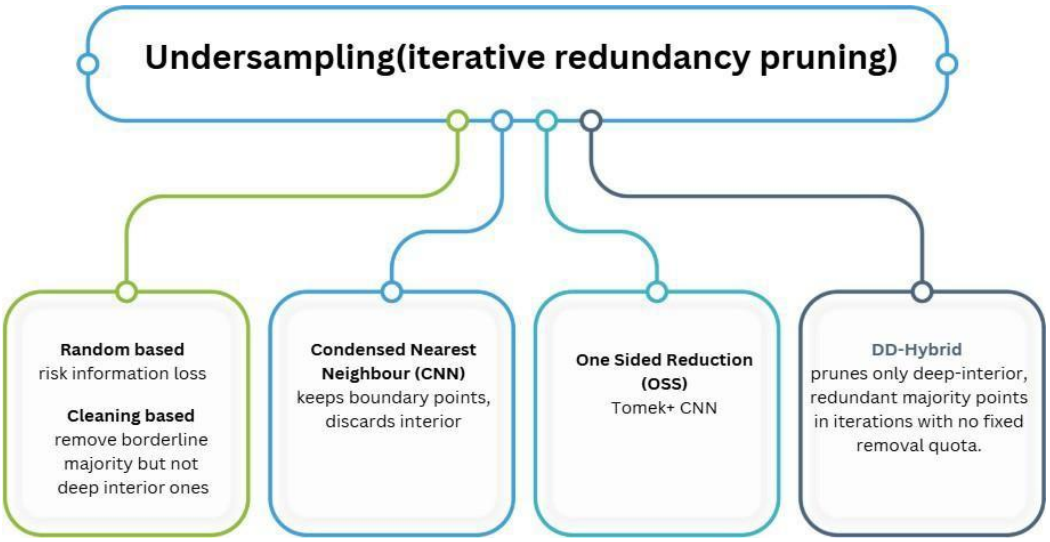


Figure 8.3: Undersampling

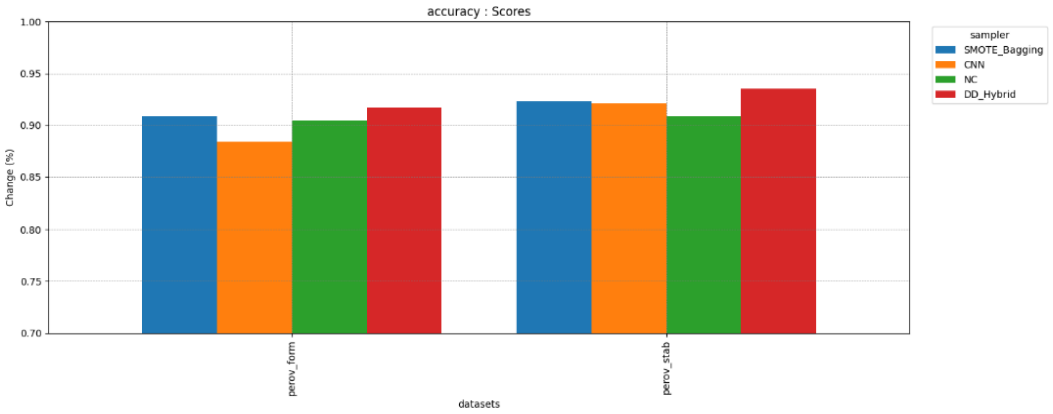
8.5 Experimental Results

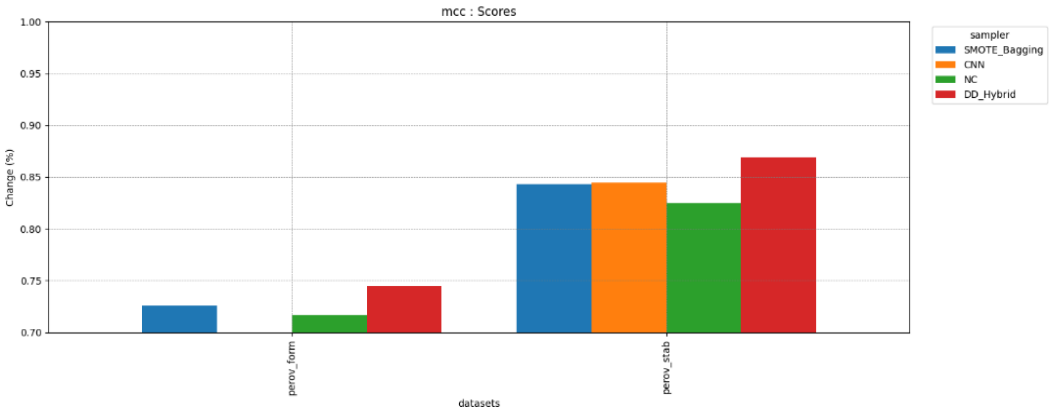
We tested DD-Hybrid on more than thirty benchmark datasets as well as on our binary perovskite datasets for formability and stability prediction. In almost all cases, DD-Hybrid performed better than SMOTE and other common sampling methods.

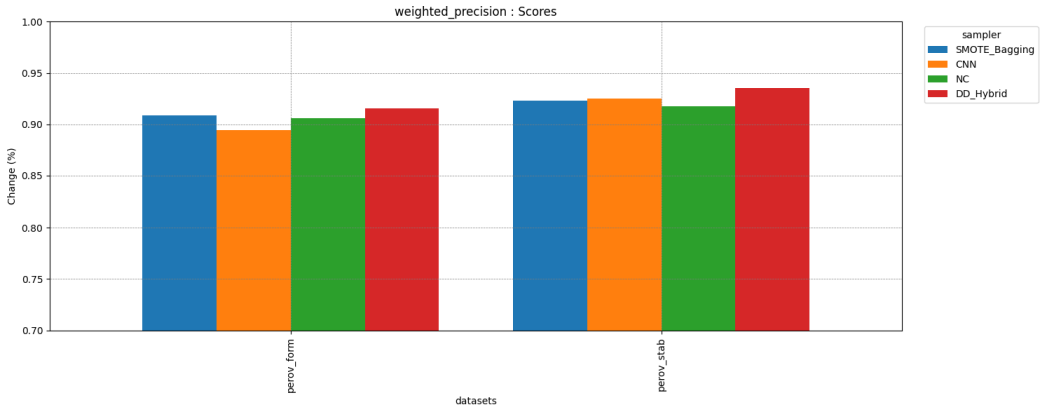
It achieved higher accuracy, weighted F1, macro precision, and weighted recall compared to SMOTE across most datasets. The improvement was more noticeable in highly imbalanced datasets, where DD-Hybrid provided cleaner decision boundaries and more stable learning.

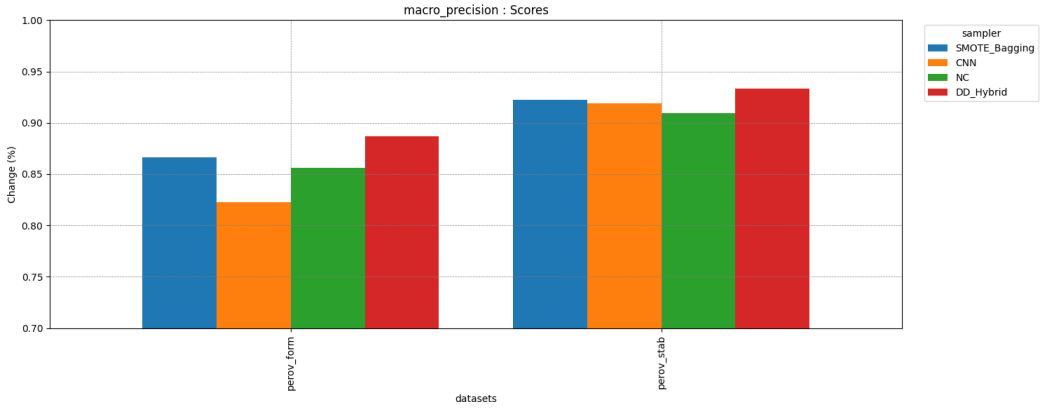
8.6 Application to Perovskite Datasets

When applied to perovskite datasets for predicting formability and stability, DD-Hybrid again showed strong results. Minority recall improved by about 11reliability increased. The balanced data helped the classifier correctly identify more stable cubic perovskites that were previously missed by other techniques.
For multiclass problems like crystal structure classification, DD-Hybrid could not be applied directly because it currently supports only binary imbalance. For those cases, we used SMOTE-Bagging, which gave the best results.


accuracy : Scores
Change (%)
1.00
0.95
0.90
0.85
0.80
0.75
0.70
perov_form
perov_stab
datasets
sampler
SMOTE_Bagging
CNN
NC
DD_Hybrid


mcc : Scores
Change (%)
1.00
0.95
0.90
0.85
0.80
0.75
0.70
perov_form
perov_stab
datasets
sampler
SMOTE_Bagging
CNN
NC
DD_Hybrid


weighted_precision : Scores
Change (%)
1.00
0.95
0.90
0.85
0.80
0.75
0.70
perov_form
perov_stab
datasets
sampler
SMOTE_Bagging
CNN
NC
DD_Hybrid


macro_precision : Scores
Change (%)
1.00
0.95
0.90
0.85
0.80
0.75
0.70
perov_form
perov_stab
datasets
sampler
SMOTE_Bagging
CNN
NC
DD_Hybrid

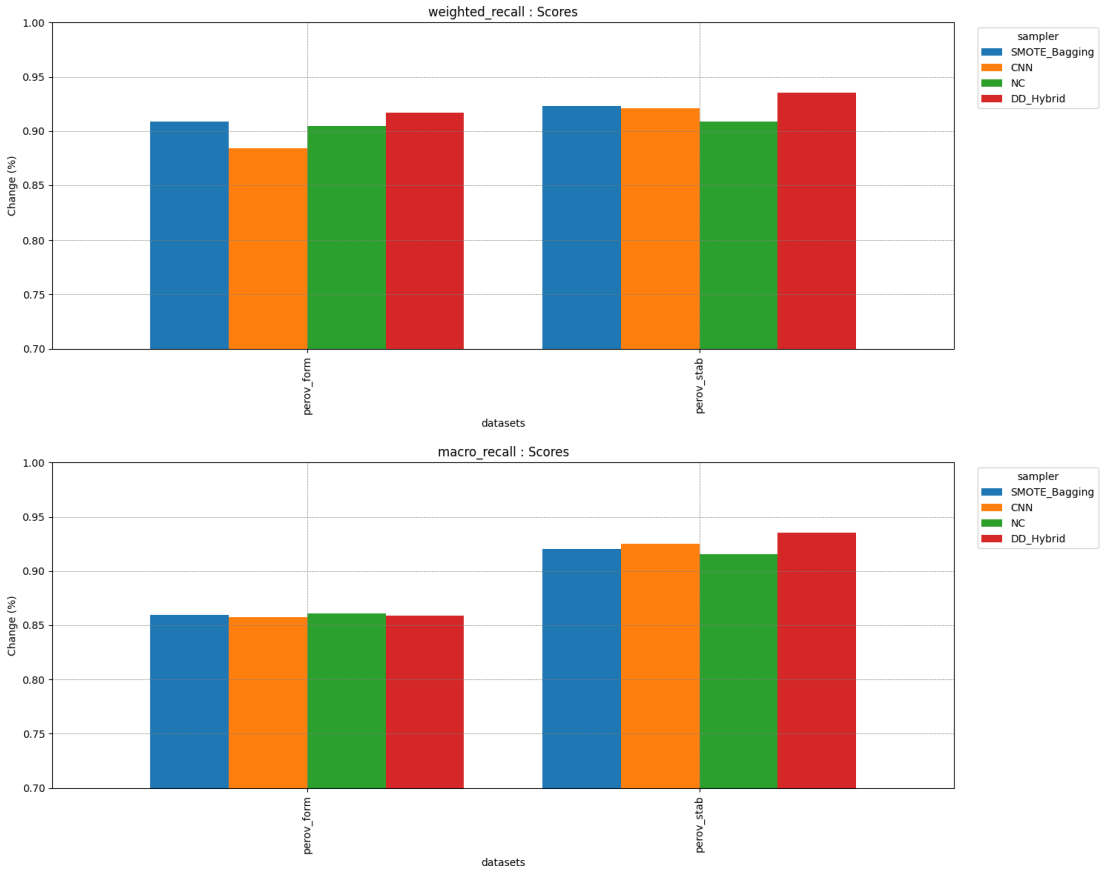


Figure 8.4: Scores for Different Sampling Methods

8.7 Limitations and Future Work

Although DD-Hybrid performed well, it has a few limitations. The k-nearest neighbor calculations can be less reliable in high-dimensional data. This can be improved using dimensionality reduction methods like PCA. When the minority class has very few samples (less than 20), the algorithm can become unstable. Also, the iterative pruning adds a bit of extra computation, though it's still manageable for datasets with up to fifty thousand samples.

In the future, DD-Hybrid can be extended to handle multiclass problems, include cost-sensitive components, or even be combined with physics-informed models for material consistency.

8.8 Summary

In this chapter, we presented our proposed method, DD-Hybrid, which we developed to handle class imbalance in materials science datasets. By combining focused oversam-

pling with iterative majority pruning, DD-Hybrid creates a balanced and clean dataset that improves the performance and reliability of machine learning models.

When tested on both benchmark and real perovskite datasets, DD-Hybrid achieved the best overall performance among all methods tested. It not only balanced the data effectively but also preserved the true structure of the dataset, making it more useful for real-world materials discovery tasks.

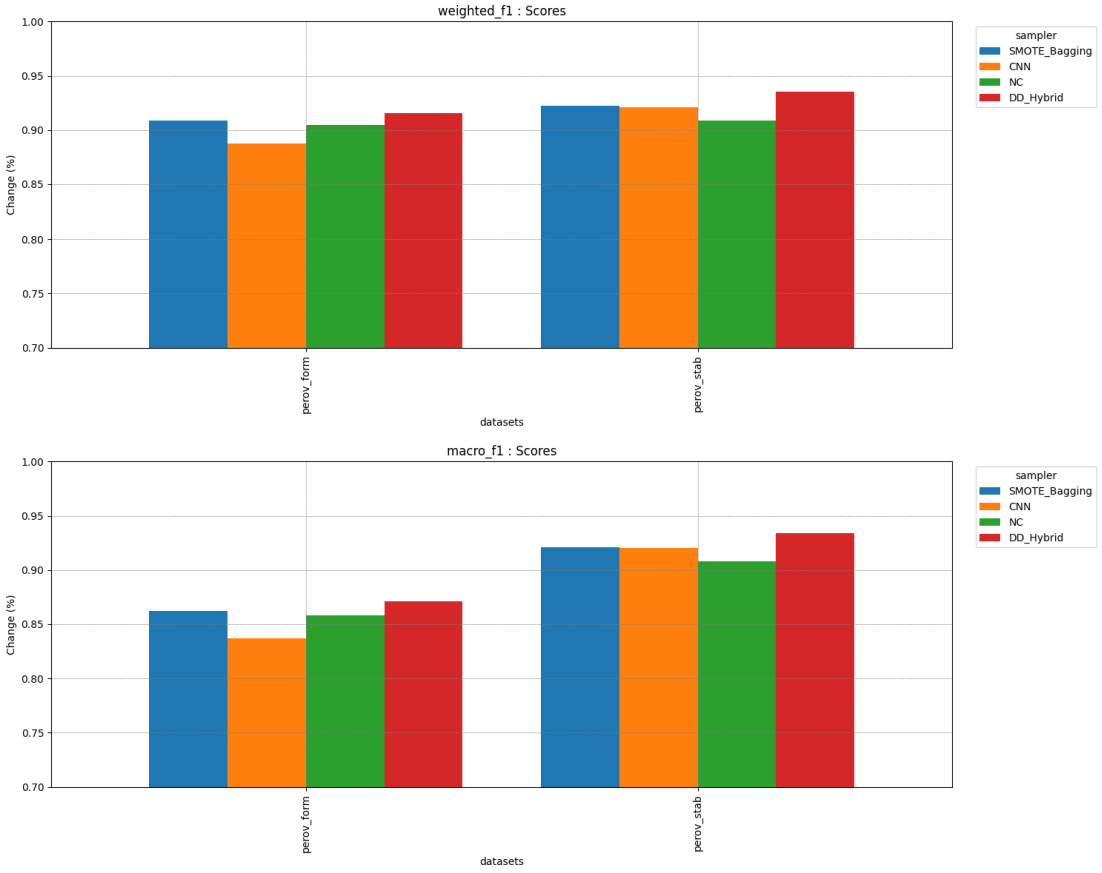


Figure 8.5: Weighted_f1 and Macro_f1 Scores for Different Sampling Methods

Chapter 9

DEMONSTRATION OF OUTCOME BASED EDUCATION (OBE)

9.1 Addressed COs and POs

COs	POs	Explanation/Justification
CO1	PO2	Our project addresses the slow and costly nature of traditional materials discovery methods, which is a significant challenge in material science and engineering.
CO2	PO4	This project synthesizes data from materials databases and applies ML models based on those needs.
CO3	PO8	Ethical consideration were taken into account when designing the experiments. None of the methods used in this project has conflict with professional norms.
CO4	PO5	We used a combination of cutting-edge software libraries and programming languages designed for data analysis and machine learning in order to adopt contemporary technical resources and techniques for the problem's solution.
CO5	PO11	To prepare a management plan and budgetary implications for the solution of the problem, we meticulously outlined the project phases, resource requirements. We established reporting schedules, meeting frequencies, and communication channels to ensure effective coordination among team members and stakeholders.
CO6	PO6	This project discusses societal benefits like eco-friendly materials and cost-effective solutions. It also reflects on reducing environmental harm and improving public safety indirectly through better materials.
CO7	PO7	We address sustainability by aiming to reduce reliance on hazardous substances and promote green material design, showing clear environmental considerations.
CO8	PO3	
CO9	PO9	Throughout the project, we emphasized the importance of working effectively both as individual contributors and as a cohesive team. Each member took responsibility for specific tasks, ensuring their timely completion and contributing to the overall project progress.
CO10	PO10	Throughout the project, we produced a number of technical reports, design documentation and powerful presentations to illustrate our approach.
CO11	PO12	

Table 9.1: CO and PO Justification.

POs	Statement	Different Aspects	Put Tick (√)
PO3	Design/development of solutions: Design solutions for complex electrical and electronic engineering problems and design systems, components or processes that meet specified needs with appropriate consideration for public health and safety, cultural, societal, and environmental considerations.	Public health	
		Safety	
		Cultural	
		Societal	
		Environmental	√
PO4	Investigation: Conduct investigations of complex electrical and electronic engineering problems using research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of information to provide valid conclusions.	Design of experiments	√
		Analysis and interpretation of data	√
		Synthesis of information	√
PO6	The engineer and society: Apply reasoning informed by contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to professional engineering practice and solutions to complex electrical and electronic engineering problems.	Societal	√
		Health	
		Safety	√
		Legal	
		Cultural	
PO7	Environment and sustainability: Understand and evaluate the sustainability and impact of professional engineering work in the solution of complex electrical and electronic engineering problems in societal and environmental contexts.	Societal	√
		Environmental	√
PO8	Ethics: Apply ethical principles embedded with religious values, professional ethics and responsibilities, and norms of electrical and electronic engineering practice.	Religious values	
		Professional ethics and responsibilities	√
		Norms	√
PO9	Individual work and teamwork: Function effectively as an individual, and as a member or leader in diverse teams and in multi-disciplinary settings.	Individual	√
		Teamwork	√
PO10	Communication: Communicate effectively on complex engineering activities with the engineering community and with society at large, such as being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.	Comprehend and write effective reports	√
		Design documentation	√
		Make effective presentations	√
		Give and receive clear instructions	√
PO11	Project management and finance: Demonstrate knowledge and understanding of engineering management principles and economic decision-making and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.	Engineering management principles	
		Economic decision-making	√
		Manage projects	√
		Multidisciplinary environments	√

Table 9.2: POs Addressed

9.2 Addressed Knowledge Profiles (K3-K8)

K	Knowledge Profile (Attribute)	Put Tick (√)
K3	A systematic, theory-based formulation of engineering fundamentals required in the engineering discipline	√
K4	Engineering specialist knowledge that provides theoretical frameworks and bodies of knowledge for the accepted practice areas in the engineering discipline; much is at the forefront of the discipline	√
K5	Knowledge that supports engineering design in a practice area	√
K6	Knowledge of engineering practice (technology) in the practice areas in the engineering discipline	√
K7	Comprehension of the role of engineering in society and identified issues in engineering practice in the discipline: ethics and the engineer's professional responsibility to public safety; the impacts of engineering activity; economic, social, cultural, environmental and sustainability	√
K8	Engagement with selected knowledge in the research literature of the discipline	√

Table 9.3: Addressed Knowledge Profiles

9.3 Addressed Attributes of Ranges of Complex Engineering Problem Solving (P1 – P7)

P	Range of Complex Engineering Problem Solving	Put Tick (√)
Attribute	Complex Engineering Problems have characteristic P1 and some or all of P2 to P7:	
Depth of knowledge required	P1: Cannot be resolved without in-depth engineering knowledge at the level of one or more of K3, K4, K5, K6 or K8 which allows a fundamentals-based, first principles analytical approach	√
Range of conflicting requirements	P2: Involve wide-ranging or conflicting technical, engineering and other issues	√
Depth of analysis required	P3: Have no obvious solution and require abstract thinking, originality in analysis to formulate suitable models	√
Familiarity of issues	P4: Involve infrequently encountered issues	√
Extent of applicable codes	P5: Are outside problems encompassed by standards and codes of practice for professional engineering	√
Extent of stakeholder involvement and conflicting requirements	P6: Involve diverse groups of stakeholders with widely varying needs	
Interdependence	P7: Are high level problems including many component parts or sub-problems	

Table 9.4: Addressed Attributes of Ranges of Complex Engineering Problem Solving.

9.4 Addressed Attributes of Ranges of Complex Engineering Activities (A1 – A5)

A	Range of Complex Engineering Activities	Put Tick (√)
Attribute	Complex activities means (engineering) activities or projects that have some or all of the following characteristics:	
Range of resources	A1: Involve the use of diverse resources (and for this purpose resources include people, money, equipment, materials, information and technologies)	√
Level of interaction	A2: Require resolution of significant problems arising from interactions between wide-ranging or conflicting technical, engineering or other issues	√
Innovation	A3: Involve creative use of engineering principles and research-based knowledge in novel ways	√
Consequences for society and the environment	A4: Have significant consequences in a range of contexts, characterized by difficulty of prediction and mitigation	√
Familiarity	A5: Can extend beyond previous experiences by applying principles-based approaches	√

Table 9.5: Addressed Attributes of Ranges of Complex Engineering Activities.

.

Chapter 10

CONCLUSION

This work set out to address one of the most persistent challenges in materials informatics — the problem of class imbalance in machine learning-based materials prediction. Using a range of public datasets on perovskite formability and stability, we examined existing sampling methods such as random oversampling, undersampling, CNN, NCL, and SMOTE-Bagging. Through systematic experimentation under a fair cross-validation setup, we observed that while traditional methods improved recall to some extent, they often produced noisy synthetic samples or lost valuable information during undersampling. These shortcomings motivated the design of our proposed Difficulty-Directed Hybrid Sampling (DD-Hybrid) method.

The proposed DD-Hybrid approach was developed to selectively balance the dataset by focusing on informative and difficult regions in the feature space. It combines targeted oversampling and iterative majority pruning to reduce bias toward the dominant class while preserving essential boundary information. Experimental evaluations showed that DD-Hybrid consistently improved classification performance, achieving higher Macro-F1 and MCC scores compared to other techniques. When applied to perovskite datasets, it significantly enhanced the prediction of stable and formable compounds, leading to more reliable discovery of promising material candidates.

Despite its success, the method has certain limitations. Its dependence on k-nearest neighbor-based neighborhood estimation makes it less effective in very high-dimensional data, and it can become unstable when the minority class has very few samples. The iterative pruning process also adds modest computational overhead. However, these trade-offs remain acceptable given the performance gains achieved across multiple datasets.

Looking ahead, several promising directions can further extend this work. Integrating DD-Hybrid with physics-informed machine learning models could ensure that generated samples remain physically meaningful. The use of generative models may enable the creation of new material compositions beyond existing datasets. Incorporating explainable AI frameworks could enhance the interpretability of predictions, helping researchers better understand why certain compounds are classified as stable or formable. Finally, adopting multi-objective optimization could balance multiple material properties simultaneously, paving the way for more comprehensive and efficient materials discovery pipelines.

In summary, this research demonstrates that by addressing class imbalance in a data-driven yet physically informed manner, we can make machine learning a more dependable tool for accelerating materials discovery. The proposed DD-Hybrid framework not only advances computational materials science but also provides a foundation for developing more adaptive, interpretable, and scalable methods in the future.

REFERENCES

1. M. U. Ahmad, A. A. Rahman Akib, M. M. Sarker Raihan, and A. B. Shams, "ABO_3 Perovskites' Formability Prediction and Crystal Structure Classification using Machine Learning," *arXiv preprint*, vol. 2022, no. 10, pp. 1–16, 2022. [Online].

   Available: https://arxiv.org/abs/2201.07831.

2. X. Zhai, F. Ding, Z. Zhao, A. Santomauro, F. Luo, and J. Tong, "Predicting the formation of fractionally doped perovskite oxides by a function-confined machine learning method," *Commun. Mater.*, vol. 3, p. 42, 2022. [Online].
   Available: https://doi.org/10.1038/s43246-022-00269-9.

3. D. Ni, W. Wu, Y. Guo, S. Gong, and Q. Wang, "Identifying key parameters for predicting materials with low defect generation efficiency by machine learning," *Comput. Mater. Sci.*, vol. 191, p. 110306, 2021. [Online].
   Available: https://doi.org/10.1016/j.commatsci.2021.110306.

4. S. Behara, T. Poonawala, and T. Thomas, "Crystal structure classification in ABO_3 perovskites via machine learning," *Comput. Mater. Sci.*, vol. 188, p. 110191, 2020. [Online].
   Available: https://doi.org/10.1016/j.commatsci.2020.110191.

5. G. S. Thoppil and A. Alankar, "Predicting the formation and stability of oxide perovskites by extracting underlying mechanisms using machine learning," *arXiv preprint*, vol. 2022, no. 10, pp. 1–20, 2022. [Online].
   Available: https://arxiv.org/abs/2201.07831.

6. S. Bhattacharya and A. Roy, "Linking stability with molecular geometries of perovskites and lanthanide richness using machine learning methods," *arXiv preprint*, vol. 2024, no. 9, pp. 1–30, 2024. [Online].
   Available: https://arxiv.org/abs/2409.13007.

7. Q. Tao, P. Xu, M. Li, and W. Lu, "Machine learning for perovskite materials design and discovery," *npj Comput. Mater.*, vol. 7, no. 23, 2021. [Online]. Available: https://doi.org/10.1038/s41524-021-00495-8.

8. M. Chen, Z. Yin, Z. Shan, X. Zheng, L. Liu, Z. Dai, J. Zhang, S. (F.) Liu, and Z. Xu, "Application of machine learning in perovskite materials and devices: A review," *J. Energy Chem.*, vol. 94, pp. 254–272, 2024. [Online]. Available:

https://doi.org/10.1016/j.jechem.2024.02.035.

9. Y. Zhang and X. Xu, "Predicting lattice parameters for orthorhombic distorted-perovskite oxides via machine learning," *Solid State Sci.*, vol. 113, p. 106541, 2021. [Online].
Available: https://doi.org/10.1016/j.solidstatesciences.2021.106541.

10. A. Talapatra, B. P. Uberuaga, C. R. Stanek, and G. Pilania, "A machine learning approach for the prediction of formability and thermodynamic stability of single and double perovskite oxides," *Chem. Mater.*, vol. 33, pp. 845–858, 2021. [Online].
Available: https://doi.org/10.1021/acs.chemmater.0c03402.

11. N. Gupta, R. Kumar, A. Alankar, et al., "Machine learning-based discovery of novel oxide and halide perovskites for energy storage," *J. Alloys Compd.*, vol. 1010, p. 177470, 2025. [Online]. Available: https://doi.org/10.1016/j.jallcom.2024.177470.

12. E. T. Chenebuah and D. T. Chenebuah, "An inorganic ABX3 perovskite materials dataset for target property prediction and classification using machine learning," *arXiv preprint*, vol. 2023, no. 7, pp. 1–15, 2023. [Online]. Available: https://github.com/chenebuah/ML_abx3_dataset.

13. A. Newaz, A. U. R. Adib, and T. Jabid, "iCost: A novel instance complexity-based cost-sensitive learning framework," *arXiv preprint*, vol. 2024, no. 10, pp. 1–20, 2024. [Online]. Available: https://doi.org/10.1016/j.knosys.2024.108634.

14. M. S. Mohosheu, A. Al-Amin, M. A. al Noman, T. Jabid, and A. Newaz, "A comprehensive evaluation of sampling techniques in addressing class imbalance across diverse datasets," *IEEE*, vol. 6th International Conference on Electrical Engineering and Information & Communication Technology, Dhaka, Bangladesh, 2024. [Online]. Available: https://doi.org/10.1109/ICEEICT.2024.177470.

15. R. Zhao, B. Xing, H. Mu, Y. Fu, and L. Zhang, "Evaluation of performance of machine learning methods in mining structure–property data of halide perovskite materials," *Chinese Phys. B*, vol. 31, p. 056302, 2022. [Online]. Available: https://doi.org/10.1088/1674-1056/ac5d2d.

16. D. Vijay Anand, Q. Xu, J. J. Wee, K. Xia, and T. C. Sum, "Topological feature engineering for machine learning based halide perovskite materials design," *npj Computational Materials*, vol. 8, p. 203, 2022. [Online]. Available: https://doi.org/10.1038/s41524-022-00883-8.

17. B. Deng, P. Zhong, K. Jun, J. Riebesell, K. Han, C. J. Bartel, and G. Ceder, "CHGNet as a pretrained universal neural network potential for charge-informed atomistic modelling," *Nature Machine Intelligence*, vol. 5, pp. 1031-1041, 2023. [Online].

Available: https://doi.org/10.1038/s42256-023-00716-3.

18. K. T. Butler, J. M. Frost, J. M. Skelton, K. L. Svanea, and A. Walsh, "Computational materials design of crystalline solids," *Chem. Soc. Rev.*, vol. 45, pp. 1826-1854, 2016. [Online]. Available: https://doi.org/10.1039/c5cs00841g.

19. X. Cai, Y. Zhang, Z. Shi, Y. Chen, Y. Xia, A. Yu, F. Xie, H. Shao, H. Zhu, D. Fu, Y. Zhan, and H. Zhang, "Discovery of lead-free perovskites for high-performance solar cells via machine learning: Ultrabroadband absorption, low radiative recombination, and enhanced thermal conductivities," *Adv. Sci.*, vol. 9, p. 2103648, 2022. [Online]. Available: https://doi.org/10.1002/advs.202103648.

20. V. Gladkikh, D. Y. Kim, A. Hajibabaei, A. Jana, C. W. Myung, and K. S. Kim, "Machine learning for predicting the band gaps of ABX3 perovskites from elemental properties," *J. Phys. Chem. C*, vol. 124, pp. 8905-8918, 2020. [Online]. Available: https://doi.org/10.1021/acs.jpcc.9b11768.

21. G. H. Gu, J. Jang, J. Noh, A. Walsh, and Y. Jung, "Perovskite synthesizability using graph neural networks," *npj Computational Materials*, vol. 8, p. 71, 2022. [Online]. Available: https://doi.org/10.1038/s41524-022-00757-z.

22. J. Im, S. Lee, T.-W. Ko, H. W. Kim, Y. Hyon, and H. Chang, "Identifying Pb-free perovskites for solar cells by machine learning," *npj Computational Materials*, vol. 5, p. 37, 2019. [Online]. Available: https://doi.org/10.1038/s41524-019-0177-0.

23. I. Kuzmanovski, S. Dimitrovska-Lazova, and S. Aleksovska, "Examination of the influence of different variables on prediction of unit cell parameters in perovskites using counter-propagation artificial neural networks," *Chem. Soc. Rev.*, vol. 45, pp. 1826-1854, 2012.

24. W. Li, R. Jacobs, and D. Morgan, "Data and supplemental information for predicting the thermodynamic stability of perovskite oxides using machine learning models," *Data in Brief*, vol. 19, pp. 261-263, 2018. [Online]. Available: https://doi.org/10.1016/j.dib.2018.05.007.

25. W. Li, R. Jacobs, and D. Morgan, "Predicting the thermodynamic stability of perovskite oxides using machine learning models," *Comput. Mater. Sci.*, vol. 150, pp. 454-463, 2018. [Online]. Available: https://doi.org/10.1016/j.commatsci.2018.04.033.

26. S. Liu, J. Wang, Z. Duan, K. Wang, W. Zhang, R. Guo, F. Xie, "Simple Structural Descriptor Obtained from Symbolic Classification for Predicting the Oxygen Vacancy Defect Formation of Perovskites," *ACS Appl. Mater. Interfaces*, vol. 14, pp. 11758-11767, 2022. [Online]. Available: https://doi.org/10.1021/acsami.1c24003.

27. Y. Priya, "Machine learning aided design of perovskite oxide materials for photocatalytic water splitting," *J. Energy Chem.*, vol. 60, pp. 351-359, 2021. [Online]. Available: https://doi.org/10.1016/j.jechem.2021.01.035.

28. S. Rath, S. P. G., N. N., and T. Thomas, "Discovery of direct band gap perovskites for light harvesting by using machine learning," *Comput. Mater. Sci.*, vol. 210, p. 111476, 2022. [Online]. Available: https://doi.org/10.1016/j.commatsci.2022.111476.

29. K. Takahashi, L. Takahashi, I. Miyazato, and Y. Tanaka, "Searching for hidden perovskite materials for photovoltaic systems by combining data science and first principle calculations," *ACS Photonics*, vol. 7, pp. 1571-1580, 2021. [Online]. Available: https://doi.org/10.1021/acsphotonics.7b01479.

30. W. Lu, Q. Tao, Y. Sheng, L. Li, and M. Li, "Machine learning aided design of perovskite oxide materials for photocatalytic water splitting," *J. Energy Chem.*, vol. 60, pp. 351-359, 2021. [Online]. Available: https://doi.org/10.1016/j.jechem.2021.01.035.

31. J. Williams, A. Mukherjee, and K. Rajan, "Deep learning based prediction of perovskite lattice parameters from Hirshfeld surface fingerprints," *J. Phys. Chem. Lett.*, vol. 11, pp. 7462-7468, 2020. [Online]. Available: https://doi.org/10.1021/jcc.21707.

32. C. W. Yap, "PaDEL-Descriptor: An open source software to calculate molecular descriptors and fingerprints," *J. Comput. Chem.*, vol. 32, pp. 1466-1474, 2011. [Online]. Available: https://doi.org/10.1002/jcc.21707.

33. T. Xie and J. C. Grossman, "Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties," *Phys. Rev. Lett.*, vol. 120, p. 145301, 2018. [Online]. Available: https://doi.org/10.1103/PhysRevLett.120.145301.

34. G. Cheng, X.-G. Gong, and W.-J. Yin, "Crystal structure prediction by combining graph network and optimization algorithm," *Nature Communications*, vol. 13, p. 1492, 2022. [Online]. Available: https://doi.org/10.1038/s41467-022-29241-4.

35. Z. Li, Q. Xu, Q. Sun, Z. Hou, and W.-J. Yin, "Thermodynamic Stability Landscape of Halide Double Perovskites via High-Throughput Computing and Machine Learning," *Adv. Funct. Mater.*, vol. 29, p. 1807280, 2019. [Online]. Available: https://doi.org/10.1002/adfm.201807280.

36. S. Zhang, T. Lu, P. Xu, Q. Tao, M. Li, and W. Lu, "Predicting the Formability of Hybrid Organic–Inorganic Perovskites via an Interpretable Machine Learning Strategy," *J. Phys. Chem. Lett.*, vol. 12, pp. 7423−7430, 2021. [Online]. Available: https://doi.org/10.1021/acs.jpclett.1c02053.

37. Q. Zhang, J. Wang, G. Liu, "Auxiliary guidance manufacture and revealing potential mechanism of perovskite solar cell using machine learning," *J. Energy Chem.*, vol. 77, pp. 200–208, 2023. [Online]. Available: https://doi.org/10.1016/j.jechem.2023.

38. J. Chen, W. Xu, and R. Zhang, "Δ-Machine learning-driven discovery of double hybrid organic–inorganic perovskites," *J. Mater. Chem. A*, vol. 10, pp. 1402−1413, 2022.

[Online]. Available: https://doi.org/10.1039/d1ta09911f.

39. P. Branco, L. Torgo, R. P. Ribeiro, "A survey of predictive modeling on imbalanced domains," *ACM Computing Surveys (CSUR)*, vol. 49, no. 2, pp. 1–50, 2016. [Online]. Available: https://doi.org/10.1145/2899478.

40. G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, G. Bing, "Learning from class-imbalanced data: Review of methods and applications," *Expert Systems with Applications*, vol. 73, pp. 220–239, 2017. [Online]. Available: https://doi.org/10.1016/j.eswa.2016.12.031.

41. A. Newaz, S. Muhtadi, F. S. Haq, "An intelligent decision support system for the accurate diagnosis of cervical cancer," *Knowledge-Based Systems*, vol. 245, p. 108634, 2022. [Online]. Available: https://doi.org/10.1016/j.knosys.2022.108634.

42. A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, F. Herrera, *Learning from imbalanced data sets*, vol. 10, Springer, 2018.

43. S. Susan, A. Kumar, "The balancing trick: Optimized sampling of imbalanced datasets—a brief survey of the recent state of the art," *Engineering Reports*, vol. 3, no. 4, p. e12298, 2021. [Online]. Available: https://doi.org/10.1002/eng2.12298.

44. M. S. Santos, P. H. Abreu, N. Japkowicz, A. Fernández, J. Santos, "A unifying view of class overlap and imbalance: Key concepts, multi-view panorama, and open avenues for research," *Information Fusion*, vol. 89, pp. 228–253, 2023. [Online]. Available: https://doi.org/10.1016/j.inffus.2023.02.008.

45. C. Elkan, "The foundations of cost-sensitive learning," in *International Joint Conference on Artificial Intelligence*, vol. 17, Lawrence Erlbaum Associates Ltd, 2001, pp. 973–978.

46. N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intelligent Data Analysis*, vol. 6, no. 5, pp. 429-449, Jan. 2002. [Online]. Available: https://doi.org/10.3233/IDA-2002-6504.

47. A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, *Learning from Imbalanced Data Sets*, Cham: Springer International Publishing, 2018. [Online]. Available: https://doi.org/10.1007/978-3-319-98074-4.

48. P. Branco, L. Torgo, and R. P. Ribeiro, "A Survey of Predictive Modeling on Imbalanced Domains," *ACM Comput. Surv.*, vol. 49, no. 2, pp. 1–50, Jun. 2017. [Online]. Available: https://doi.org/10.1145/2907070.

49. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16,
pp. 321–357, 2002. [Online]. Available: https://doi.org/10.1613/JAIR.953.

50. D. Elreedy and A. F. Atiya, "A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance," *Inf. Sci. (N.Y.)*, vol. 505,
pp. 32–64, Dec. 2019. [Online]. Available: https://doi.org/10.1016/j.ins.2019.07.070.

51. L. Ma and S. Fan, "CURE-SMOTE algorithm and hybrid algorithm for feature selection and parameter optimization based on random forests," *BMC Bioinformatics*, vol. 18, no. 1, p. 169, Dec. 2017. [Online]. Available: https://doi.org/10.1186/s12859-017-1578-z.

52. G. Douzas and F. Bacao, "Geometric SMOTE a geometrically enhanced drop-in replacement for SMOTE," *Inf. Sci. (N.Y.)*, vol. 501, pp. 118–135, Oct. 2019. [Online]. Available: https://doi.org/10.1016/j.ins.2019.06.007.

16% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Match Groups

151 Not Cited or Quoted 13%
Matches with neither in-text citation nor quotation marks

6 Missing Quotations 1%
Matches that are still very similar to source material

18 Missing Citation 2%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

10% Internet sources

9% Publications

11% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.