University “St. Kliment Ohridski”
Bitola
Faculty of Information and
Communication Technology - Bitola
Republic of North Macedonia

PROCEEDINGS

15th International Conference on APPLIED INTERNET AND INFORMATION TECHNOLOGIES

AIIT 2025

AIIT
International Conference

Bitola, November 7, 2025

University “St. Kliment Ohridski” Bitola
Faculty of Information and Communication Technology - Bitola
Republic of North Macedonia

PROCEEDINGS
15th International Conference on
APPLIED INTERNET AND INFORMATION TECHNOLOGIES

AIIT 2025

November 7, 2025 Bitola

Proceedings publisher and organizer of the conference:
University "St. Kliment Ohridski", Bitola, Faculty of Information and Communication Technology – Bitola, Republic of North Macedonia

For publisher:
Blagoj Ristevski, PhD, Full Professor, Dean of the Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia

Proceedings editors:
Kostandina Veljanovska, PhD
Željko Stojanov, PhD

Conference Chairmans:
Blagoj Ristevski, University "St. Kliment Ohridski", Faculty of Information and Communication Technologies, Bitola, Republic of N. Macedonia – chair
Kostandina Veljanovska, University "St. Kliment Ohridski", Faculty of Information and Communication Technologies, Bitola, Republic of N. Macedonia – co-chair
Željko Stojanov, University of Novi Sad, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia – co-chair

Technical preparation of the proceedings:

Kostandina Veljanovska, PhD
Marija Apostoloska Kondoska, MSc
Darko Pajkovski, MSc

Cover design:
Kostandina Veljanovska, PhD
Hristina Dimova Popovska, MSc

e-Proceedings
ISBN 978-608-5003-06-8

CIP - Каталогизација во публикација

CIP - Каталогизација во публикација
Национална и универзитетска библиотека "Св. Климент Охридски", Скопје

004-049.8(062)

INTERNATIONAL conference on applied internet and information technologies AIIT 2025 (15 ; 2025 ; Bitola, Republic of North Macedonia)
Proceedings / 15th International conference on applied internet and information technologies AIIT 2025, November 7 2025, Bitola, Republic of North Macedonia ; [editors Kostandina Veljanovska, Zeljko Stojanov]. - Bitola : University "St. Kliment Ohridski", Bitola Faculty of information and communication technologies, 2025. - 477 стр. : илустр. ; 30 см
Библиографија кон трудовите
ISBN 978-608-5003-06-8

а) Информатичка технологија -- Примена -- Собири
COBISS.MK-ID 67608325

Introduction

As organizing partners of 15th International Conference on Applied Internet and Information Technologies AIIT 2025, we warmly welcome all participants, researchers, and colleagues joining us from various countries and universities, united by our shared commitment to advancing knowledge in the fields of computer science, applied Internet, and information technologies.

The AIIT conference has become a long-standing tradition of excellence and collaboration, co-organized by the Faculty of Information and Communication Technologies – Bitola, University "St. Kliment Ohridski," and the Technical Faculty "Mihajlo Pupin" – Zrenjanin, University of Novi Sad, Serbia. Over the past fifteen years, this partnership has fostered not only strong academic cooperation but also genuine friendship among our institutions and scholars.

This year's conference proudly continues that tradition, bringing together innovative research, diverse perspectives, and new insights into technologies that are shaping our digital future. The Scientific Program Committee once again faced the demanding task of selecting the highest-quality papers from more than sixty submissions spanning a wide range of topics—including Artificial Intelligence, Immersive Technologies, Mathematical Simulations, Data Science and Big Data Analytics, Knowledge and IT Management, Cybersecurity, Software Engineering, Data Mining, Digital Transformation, Behavioral Economics and Business, Social Engineering, Digital Humanities, Augmented Humanity, and Hybrid Intelligence. This ensures that the program reflects both scientific rigor and creative originality.

We would like to express our sincere gratitude to all reviewers for their dedicated work, as well as to the members of the Organizing Committee for their professionalism, commitment, and enthusiasm in preparing this event.

We are confident that these proceedings will provide an enriching and thought-provoking reading experience.

Conference chairs:

Blagoj Ristevski, University "St. Kliment Ohridski", Faculty of Information and Communication Technologies, Bitola, Republic of N. Macedonia (chair)
Kostandina Veljanovska, University "St. Kliment Ohridski", Faculty of Information and Communication Technologies, Bitola, Republic of N. Macedonia (co - chair)
Željko Stojanov, University of Novi Sad, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia (co – chair)

MAIN ORGANIZERS:

Faculty of Information and Communication Technologies - Bitola
University "St. Kliment Ohridski" University - Bitola
NORTH MACEDONIA
http://fikt.uklo.edu.mk/

Technical Faculty "Mihajlo Pupin" Zrenjanin
University of Novi Sad SERBIA
http://www.tfzr.uns.ac.rs/

ORGANIZATION PARTNERS:

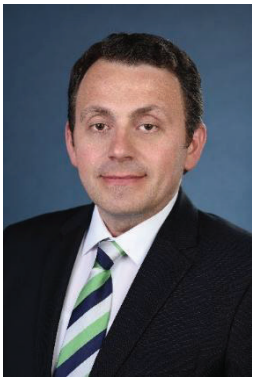
Faculty of Computer Science
Irkutsk National Research Technical University
Institute of Informational Technologies and Data Analysis
Irkutsk, RUSSIA
http://www.istu.edu/

Matrosov Institute for System Dynamics and Control Theory of Siberian Branch of Russian Academy of Sciences, Irkutsk, RUSSIA
http://idstu.irk.ru/

Irkutsk State Transport University (IrGUPS)
Irkutsk, RUSSIA
https://www.irgups.ru/

Faculty of Engineering South-west
University "Neophyte Rilsky"-Blagoevgrad
BULGARIA
http://www.swu.bg/

Conference Chairs



Blagoj Ristevski, University "St. Kliment Ohridski", Faculty of Information and Communication Technologies, Bitola, Republic of N. Macedonia (chair)

Prof. Dr. Blagoj Ristevski is a Full Professor at the Faculty of Information and Communication Technologies (FICT) at the University "St. Kliment Ohridski" - Bitola, where he currently serves as Dean. He holds a PhD in Technical Sciences from the Faculty of Electrical Engineering and Information Technologies, Institute of Computer Science and Informatics, at Ss. Cyril and Methodius University in Skopje. His research interests span Databases, Data Science, Data Mining, Big Data Analytics, Bioinformatics, Computer Graphics, and Cybersecurity. Prof. Ristevski has supervised numerous BSc, MSc, and PhD theses and has led several international research projects. He has served on the management committees of multiple COST actions, reviewed for numerous high-impact journals, and evaluated project proposals for the Horizon 2020 and Horizon Europe programs. Prof. Ristevski is also a senior member of IEEE.

Kostandina Veljanovska, University "St. Kliment Ohridski", Faculty of Information and Communication Technologies, Bitola, Republic of N. Macedonia (co – chair)

Kostandina Veljanovska, Ph.D. completed her education at the University "Sts. Kiril i Metodi", Skopje (BSc in Computer Science), at the University of Toronto, Toronto (MASc in Applied Engineering) and got her MSc and also her PhD in Technical Sciences at the University "St. Kliment Ohridski", Bitola, R. Macedonia. She has completed postdoc in Artificial Intelligence at the Laboratory of Informatics, Robotics and Microelectronics at the University of Montpellier, Montpellier, France. She worked as a Research assistant at the Faculty of Applied Science, University of Toronto, Canada. She also, worked at research team for Constraints, Learning and Agents at LIRMM, University of Montpellier. Currently, she works as a Full Professor in Artificial Intelligence and Systems, Computer Science and Computer Engineering at the Faculty of Information and Communication Technologies, University "St. Kliment Ohridski" - Bitola and serves as a Vice-dean for Science and Collaboration. Her research work is focused on artificial intelligence, machine learning techniques, intelligent systems and human - computer interaction. She participated in several international and domestic scientific projects. She has published numerous scientific papers in the area of interest, as well as several monographic items. She is a reviewing referee for well-known publishing house, journals with significant impact factor in science and also, member of editorial board of several international conferences.

Željko Stojanov, University of Novi Sad, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia (co – chair)

Željko Stojanov, Ph.D. received PhD degree in Computer science and applied informatics at University of Novi Sad, Serbia. He works as a full professor at University of Novi Sad, Technical Faculty "Mihajlo Pupin" Zrenjanin, Serbia. His research interests are in the fields of software engineering, software architecture, software life cycle, business informatics, learning and knowledge management, engineering education, and human aspects of software engineering. He is author of scientific papers published in refereed journals and in the proceedings of international conferences. He participated in several research and industrial projects at national and international levels. He has over fifteen years of experience working with small software companies as a consultant in the fields of software development, software maintenance and software process improvement.

Organizing Committee

Chairs

Kostandina Veljanovska (President), Faculty of Information and Communication Technologies - Bitola, N. Macedonia
Željko Stojanov, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia (vice-president)

Members

Blagoj Ristevski, Faculty of Information and Communication Technologies - Bitola, N. Macedonia
Ivana Berković, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Višnja Ognjenović, Technical Faculty "Mihajlo Pupin" Zrenjanin, Serbia
Eleonora Brtka, Technical Faculty "Mihajlo Pupin" Zrenjanin, Serbia
Dalibor Dobrilovic, Technical Faculty "Mihajlo Pupin" Zrenjanin, Serbia
Dragica Radosav, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Biljana Radulovic, Technical Faculty "Mihajlo Pupin" Zrenjanin, Serbia
Božidar Milenkovski, Faculty of Information and Communication Technologies - Bitola, N. Macedonia
Ljubica Kazi, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Snežana Savoska, Faculty of Information and Communication Technologies - Bitola,N. Macedonia
Vladimir Brtka, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Zoltan Kazi, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Siniša Mihajlović, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Velibor Premčevski, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Nikola Rendevski , Faculty of Information and Communication Technologies - Bitola, N. Macedonia
Aleksandra Stojkov, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Maja Gaborov, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Milica Mazalica, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Igor Vecštejn, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Marko Blažić, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Vuk Amižić, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Natasa Blazeska-Tabakovska, Faculty of Information and Communication Technologies - Bitola, N. Macedonia
Jovana Borovina, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Dalibor Šeljmeši, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Mimoza Bogdanoska Jovanovska, Faculty of Information and Communication Technologies - Bitola, N. Macedonia
Vladimir Šinik, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Nadežda Ljubojev, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Marina Blažeković Toshevski, Faculty of Information and Communication Technologies - Bitola, N. Macedonia
Hristina Dimova Popovska, Faculty of Information and Communication Technologies, University "St.Kliment Ohridski" - Bitola, N. Macedonia
Darko Pajkovski, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski" - Bitola, N. Macedonia
Marija Apostoloska Kondoska, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski" - Bitola, N. Macedonia
Milcho Prisagjanec, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski" - Bitola, N. Macedonia
Ilche Dimovski, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski" - Bitola, N. Macedonia
Zoran Pavlovski, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski" - Bitola, N. Macedonia
Vladimir Karuović, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Evgeny Cherkashin, Institute of System Dynamic and Control Theory SB RAS, Russia
Anastasia Popova, Matrosov Institute for System Dynamics and Control Theory of the Siberian Branch of the Russian Academy of Sciences, Russia
Filip Tsvetanov, South-west University "Neophyte Rilsky", Faculty of Engineering, Blagoevgrad, Bulgaria

Program Committee

Blagoj Ristevski (president), Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia,

Željko Stojanov, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia (vice-president)

Kostandina Veljanovska, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia (vice-president)
Eleonora Brtka, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Višnja Ognjenović, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Dalibor Dobrilović, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Ljubica Kazi, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Dragica Radosav, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Dragana Glušac, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Borislav Odadžić, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Miodrag Ivković, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Biljana Radulović, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Ivana Berković, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Vladimir Brtka, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Zoltan Kazi, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Jelena Stojanov, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Vesna Makitan, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Nadežda Ljubojev, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Vladimir Šinik, Technical faculty "Mihajlo Pupin", Zrenjanin, Serbia
Igor Nedelkovski, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia
Aleksandar Markoski, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia
Violeta Manevska, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia
Pece Mitrevski, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia
Ilija Jolevski, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia
Dragan Grueski, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia
Monika Markovska, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia
Snežana Savoska, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia
Sonja Mančevska, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia
Mimoza Bogdanoska Jovanovska, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia
Nataša Blažeska Tabakovska, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia
Božidar Milenkovski, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia
Zoran Kotevski, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski", Bitola, North Macedonia
Nikola Rendevski, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski",

CONTENT

AI-Based Prediction of Elastic Properties in Crystals with Class Balancing

Nora Pireci Sejdiu[1], Nikola Rendevski[2] and Blagoj Ristevski[3]

[1, 2, 3] *Faculty of Information and Communication Technologies – Bitola, University "St. Kliment Ohridski" - Bitola, Republic of Macedonia*

pireci.nora@uklo.edu.mk, nikola.rendevski@uklo.edu.mk, blagoj.ristevski@uklo.edu.mk

Abstract:

For the design of new functional materials for industrial applications, predicting the mechanical properties of materials is essential. In this study, we present an AI-powered classification framework for predicting the elastic moduli of crystalline materials based on the elastic_tensor_2015 dataset provided by the Materials Project, a publicly available repository that provides computed properties of a wide range of crystalline materials. The biggest challenge of this dataset is class imbalance, which can bias the predictive models and limit generalisation. To address this issue, we applied two oversampling techniques - ADASYN and SMOTE-Tomek - to generate synthetic minority samples and improve model learning. We trained and evaluated eight machine learning algorithms, including Balanced Random Forest, XGBoost, CatBoost, Multi-Layer Perceptron (MLP), Logistic Regression, K-Nearest Neighbors, Support Vector Machine, and Naive Bayes. Model evaluation was performed using standard metrics such as precision, recall, F1 score, and ROC-AUC, along with feature importance analysis to interpret model decisions. The results show that ensemble-based models and neural networks achieved the highest predictive performance, while simpler models such as SVM and Naive Bayes showed limited effectiveness. This study highlights the impact of data balancing and algorithm selection on machine learning models for materials informatics and provides insights into the development of accurate, data-driven tools to accelerate material discovery.

Keywords:

AI in materials science, machine learning, oversampling techniques, materials informatics, class imbalance, SMOTE-Tomek, ADASYN

1. Introduction

The discovery of new materials is one of the main pillars of technological development in various sectors. Among the many properties of materials, elastic properties represent fundamental mechanical descriptors that determine how a material responds to loads, the stability of the structure of that material and how suitable it is for specific applications. The determination of the density of these properties has been achieved through time-consuming experimental methods. This obstacle has been overcome with the help of artificial intelligence (AI) and machine learning (ML) by accelerating this process [2, 5, 11]. The practical application of ML in this field still faces various data-related challenges, which can negatively affect the reliability and generalisation ability of models. One of these challenges is class imbalance [7], [11]. In classification tasks – such as identifying materials with high or low elastic modulus- the distribution of classes is usually asymmetric. This causes ML models to bias towards the majority class, achieving a falsely high accuracy by ignoring the minority class, which often contains the most scientifically interesting or technologically valuable materials, such as materials with rare properties [10]. This problem is not limited to materials science but is present in many other scientific fields. Different authors have developed different techniques to mitigate this problem, ranging from data-level methods such as the Synthetic Minority Over-sampling Technique (SMOTE) and its variants (e.g., ADASYN, SMOTE-Tomek) to algorithmic-level methods [7], [10].

To address these gaps, this paper presents a framework for predicting elastic properties in crystals, with a particular focus on the problem of class imbalance. The main goals and contributions of the paper are:

Implementation and evaluation of two advanced oversampling techniques, SMOTE-Tomek and ADASYN, for dealing with class imbalance in Poisson ratio prediction using the elastic_tensor_2015 dataset of the Materials Project.

Extensive comparison of eight machine learning algorithms – including ensemble methods (Balanced Random Forest, XGBoost, CatBoost), neural networks (MLP), linear models (Logistic Regression, SVM), and probabilistic methods (Naive Bayes, KNN) – to identify the most robust model for this task.

Rigorous evaluation of the performance of the models through a set of metrics adapted for unbalanced data (Precision, Recall, F1-Score, ROC-AUC) and comparative analysis between rebalancing techniques.

The remainder of this paper is structured as follows. Section 2 provides an overview of the existing literature on ML in materials science and learning from uncalibrated data. Section 3 describes the methodology, Section 4 discusses the main results, while Section 5 summarises the findings and offers directions for future research.

2. Related work

The rapid spread of ML in materials science has transformed the prediction and discovery of properties.

A model that predicts mechanical/elastic behaviour is a key enabler for the selection of candidate materials. However, reliable models remain limited by data quality and distribution issues that bias performance [2,5].

A major challenge in all scientific ML has been class imbalance, where rare classes are underrepresented. Research in chemistry and molecular sciences synthesises best practices for data generation and balancing using models such as SMOTE and ADASYN [7]. Systematic studies demonstrate the importance and effectiveness of SMOTE-based oversampling and hybrid variants such as SMOTE-Tomek/SMOTE-ENN in all fields [10]. In particular, in materials science, class imbalance represents a typical example of situations with few data for the minority class. For this reason, generating additional data through synthetic samples is considered the best solution to address this problem [11].

Many studies show that actively handling data imbalance is often the factor that determines whether a model is actually useful or unreliable. In the field of structural health monitoring, the authors in [1] combined finite element simulations, signal-to-image conversion, and the SMOTE technique with a CNN to detect delamination, achieving high accuracy even when the data were sparse or distorted.

In bioinformatics, the authors in [3] reduced noise in features through empirical mode decomposition and applied SMOTE to balance the data, testing eight well-known algorithms (RF, XGBoost, LightGBM, SVM, KNN, NB, LR, MLP). They proved that ensembles of trees and gradient boosting methods give more stable results after balancing.

In geoscience, the authors in [4] analysed an environment with extreme data imbalance, comparing oversampling, undersampling, and balanced ensembles (Balanced RF, RUSBoost). They showed that unbalanced models are not practically valid, while balanced ones manage to recover signals of geological importance. They emphasised the need for spatial validation to avoid optimistic separations.

These studies form a clear methodological model: data balancing, use of ensembles and metrics suitable for unbalanced data, an approach that finds application in all scientific fields of classification. In contemporary materials science, model accuracy alone is no longer enough. The emphasis is increasingly placed on transparency and interpretability to derive clear rules related to physical principles.

The authors in [6] used interpretable machine learning models to analyse stress concentration in magnesium, applying methods such as SHapley Additive exPlanations (SHAP), Partial Dependence Plots (PDP) and feature importance analysis to explain how the models make inferences. This made it possible to link predictions to mechanistic implications and increased the reliability of the results.

In the study of perovskite optoelectronics, the authors in [9] used similar analyses to identify the most important descriptors affecting performance, using tree-based interpretation methods and a model-agnostic approach [9].

Synthesis studies provide the methodological basis for selecting techniques in data-imbalanced environments. The authors in [7] (Chemical Science) propose tools at three levels: data level (SMOTE/ADASYN families), algorithm level (class weights, cost-sensitive loss), and hybrid

approaches, and codify evaluation standards (accuracy/recall/F1, MCC, ROC-AUC/PR-AUC) for more realistic evaluation.

The authors in [10] highlight SMOTE and its hybrid variants (e.g., SMOTE-Tomek) as the most widely used approaches, while the authors in [11] underline the importance of oversampling to expand the data when minority classes represent rare but valuable behaviours.

In [8], the authors demonstrate that adaptive approaches such as ADASYN and combined approaches such as SMOTE-Tomek achieve a better balance between data generation and cleaning.

In line with the existing literature, this study contributes in three main directions:

1. Problem approach. Unlike most works in materials informatics that focus on deep architectures or balanced regression models [2,5,11], we explicitly address the problem of classifying elastic moduli under strain. To this end, we use the Materials Project dataset – elastic_tensor_2015 – considering imbalance as a fundamental modelling challenge.
2. Balancing strategies. We systematically compare two approaches to dealing with imbalance at the data level:
   - ADASYN, which creates synthetic samples near minority class hardpoints;
   - SMOTE-Tomek, which combines generation with cleaning to improve separation limits.

   These approaches reflect the latest guidelines of methodological practices in dealing with unbalanced data [7,8,10,11].
3. Model comparison and interpretability. A wide range of algorithms are tested, such as balanced ensembles (Balanced Random Forest), gradient boosting methods (XGBoost, CatBoost), neural networks (MLP), and classical models (LR, KNN, SVM, NB).

The evaluation is based on imbalance-sensitive metrics (precision, recall, F1, ROC-AUC) [3,4,7,10,12]. In addition to performance, the importance of features is analysed to identify the most influential descriptors of elastic behaviour, linking the model interpretation to the physical knowledge of materials [6,9].

3. Methodology

3.1.1. Data Source and Preprocessing

The data were taken from the Materials Project [13], elastic_tensor_2015 collection, which contains elastic properties and crystal structures of 1,181 materials.

To avoid data leakage, all previously calculated elastic properties were excluded. The features were constructed from composition and structure data using the Magpie (Materials-Agnostic Platform for Informatics and Exploration) descriptor set implemented in matminer's ElementProperty featurizer, which generates 132 features from statistical summaries of elemental properties. These were combined with structural features (volume, density, symmetry) and basic crystallographic data, totalling 141 features.

3.2. Problem Formulation and Class Imbalance Handling

The task was formulated as a binary classification to predict whether a material has a high Poisson ratio. The values were divided into two classes according to the threshold $\nu = 0.3$:

- Class 0 (majority): $\nu \leq 0.3$ (663 cases, 56.1%).
- Class 1 (minority): $\nu > 0.3$ (518 cases, 43.9%)

To mitigate the imbalance (~1.28:1), two methods were applied to the training set:

- SMOTE-Tomek: combines synthetic oversampling with cleaning of overlapping samples.
- ADASYN: generates more samples for minority cases that are harder to learn

3.3. Machine Learning Models

To provide a comparative analysis, eight algorithms representing different methods were tested:

- Balanced Random Forest, XGBoost, and CatBoost for ensembles.
- Multilayer Perceptron (MLP) for neural networks.
- Logistic Regression and SVM for linear models.
- Naive Bayes KNN for proximity and probability-based methods.

3.4. Model Training and Evaluation

The dataset was split into 70% training and 30% testing. Resampling was applied only to the training data to avoid data leakage.

The models were trained with random_state = 42, and the performance was evaluated through the metrics accuracy, precision, recall, F1-score, and ROC-AUC.

The most successful model was further analysed for the importance of features, to physically interpret the factors that affect the elastic properties of materials.

4. Results and discussion

This section presents an analysis of the experimental results, evaluating the effectiveness of different ML models and resampling techniques for predicting the elastic properties of crystalline materials. The performance is discussed with a focus on the impact of class balancing and the comparative strengths of the algorithms.

4.1. Dataset Characteristics and Class

The dataset comprised 1,181 crystalline materials with a moderate class imbalance in the target variable, is_high_poisson_ratio. The distribution showed 663 instances (56.1%) with low Poisson's ratio (False) and 518 instances (43.9%) with high Poisson's ratio (True), resulting in an imbalance ratio of 1.28:1. After the train-test split, the training set contained 464 majority class and 362 minority class instances, while the test set maintained proportional distribution. While this represents a moderate imbalance rather than extreme cases, it is sufficient to introduce bias in machine learning models if not properly addressed [7, 10].

4.2. Overall Model Performance and the Impact of Resampling

The application of SMOTE-Tomek and ADASYN resampling techniques successfully balanced the training data, with SMOTE-Tomek creating perfectly balanced classes (410 instances each) and ADASYN achieving near-balance (464 majority vs. 426 minority). This balancing proved crucial for equitable learning across both classes.

A key observation is the strong performance of tree-based ensemble methods across both resampling techniques. XGBoost achieved the highest F1-score (0.789) with ADASYN, while Balanced Random Forest and CatBoost consistently delivered robust performance with F1-scores above 0.760. This indicates that these ensemble methods effectively learned the complex relationships between composition/structure features and elastic behaviour.

The performance levels (F1-scores: 0.697-0.789) represent realistic and scientifically meaningful results, demonstrating that predicting Poisson's ratio from fundamental material characteristics is a challenging but feasible task. These results contrast with the previously reported near-perfect scores that were artefacts of data leakage from using pre-computed elastic properties as features.

Conversely, Naive Bayes and K-Nearest Neighbors exhibited relatively weaker performance. For Naive Bayes, this is likely due to the violation of its core assumption of feature independence, which is often not the case for correlated materials descriptors. The SVM showed moderate performance (F1-scores: 0.732-0.736), suggesting it struggled with the high-dimensional feature space.

Table 1:

Model performance comparison across resampling techniques

Model	Sampling technique	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	SMOTE-Tomek	0.792	0.747	0.795	0.770	0.865
	ADASYN	0.789	0.745	0.788	0.766	0.853
CatBoost	SMOTE-Tomek	0.808	0.765	0.814	0.789	0.863
	ADASYN	0.769	0.720	0.776	0.747	0.858

XGBoost	SMOTE-Tomek	0.794	0.758	0.782	0.770	0.865
	ADASYN	0.789	0.755	0.769	0.762	0.858
Balanced Random Forest	SMOTE-Tomek	0.746	0.688	0.776	0.729	0.828
	ADASYN	0.741	0.705	0.705	0.705	0.826
MLP	SMOTE-Tomek	0.721	0.656	0.769	0.708	0.805
	ADASYN	0.744	0.702	0.724	0.713	0.807
K-Nearest Neighbors	SMOTE-Tomek	0.665	0.578	0.878	0.697	0.792
	ADASYN	0.701	0.620	0.827	0.709	0.787
Naive Bayes	SMOTE-Tomek	0.749	0.722	0.699	0.710	0.826
	ADASYN	0.780	0.757	0.737	0.747	0.857
Support Vector Machine	SMOTE-Tomek	0.752	0.698	0.769	0.732	0.859
	ADASYN	0.763	0.722	0.750	0.736	0.846

4.3. Comparative Analysis of SMOTE-Tomek and ADASYN

Both resampling techniques demonstrated effectiveness, but ADASYN provided a consistent advantage for the top-performing ensemble methods:

ADASYN Superiority: The three best models (XGBoost, Balanced Random Forest, and CatBoost) all achieved their highest F1-scores with ADASYN. XGBoost showed particularly strong improvement with ADASYN, increasing its F1-score from 0.747 to 0.789.

SMOTE-Tomek Consistency: While ADASYN produced the peak performances, SMOTE-Tomek demonstrated more consistent results across different model types. The "cleaning" step provided by Tomek Links may have created more stable decision boundaries.

Model-Specific Responses: The MLP showed better performance with SMOTE-Tomek (F1: 0.747 vs 0.710), suggesting that the cleaner boundaries from Tomek Links benefited the neural network. Conversely, distance-based methods like KNN showed minimal difference between techniques.

The adaptive nature of ADASYN, which focuses synthetic sample generation on "hard-to-learn" minority instances, appears particularly well-suited for ensemble methods that can leverage these challenging examples to build more robust decision boundaries.

4.4. ROC-AUC Analysis

Figure 1 shows the ROC curves for all models using SMOTE-Tomek, while Figure 2 shows the results when ADASYN was used. These graphs help to understand the differences between materials with high and low Poisson ratios.

The ensemble methods showed high discrimination ability. The Balanced Random Forest model and CatBoost with ADASYN resampling achieved the highest AUC value (0.865), which means that they achieve a good balance between accurate classifications and low errors. This shows that ensemble models, which combine multiple decision trees, are effective for this type of problem in materials science.

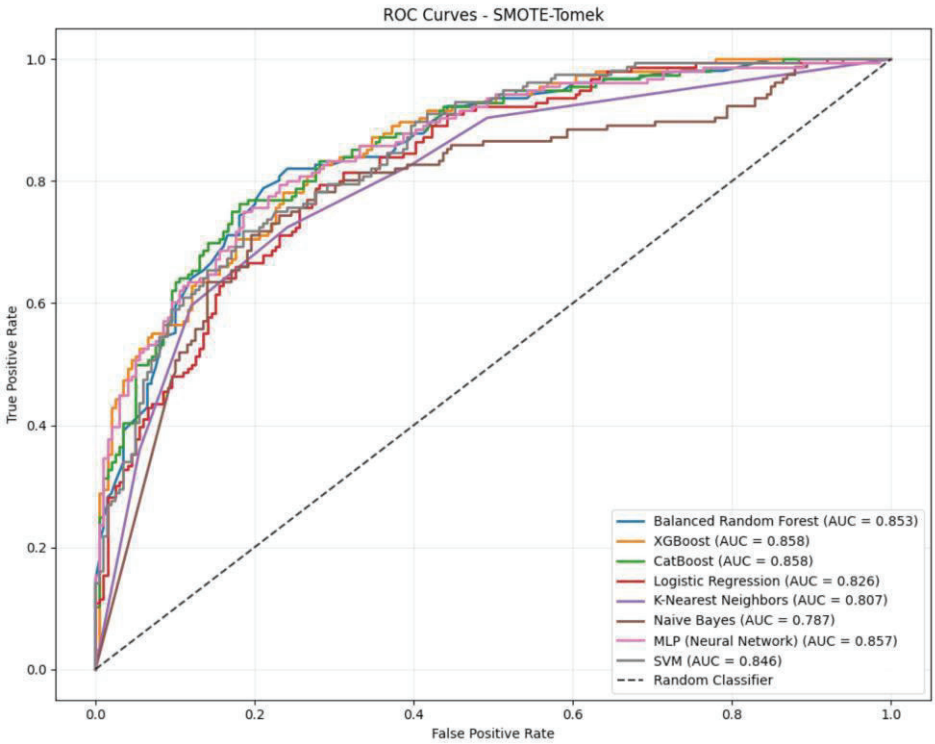


Figure 1: ROC-AUC analysis of ML models using SMOTE-Tomek for class balancing

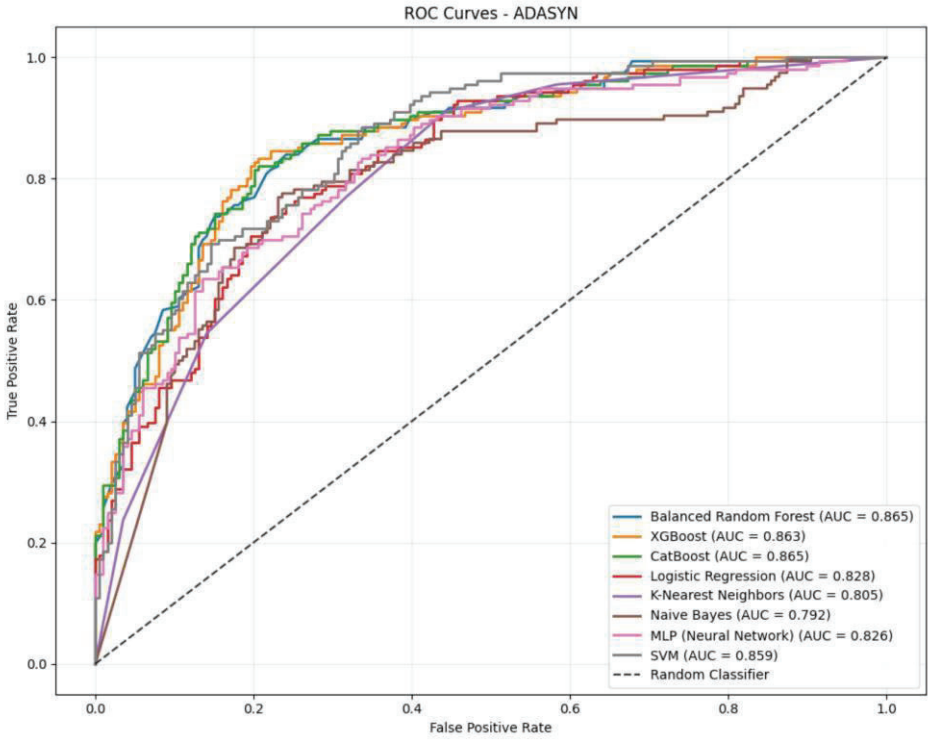


Figure 2: ROC-AUC analysis of ML models using ADASYN for class balancing

The comparison between the two resampling techniques shows that ADASYN generally gave better results for ensembles, while SMOTE-Tomek provided more consistent results across all models.

The MLP (neural network) model had consistent performance with both methods (AUC around 0.85), indicating that it is not greatly affected by the choice of the resampling technique.

The performance hierarchy shown in the figure is consistent with the other evaluation metrics:

Balanced Random Forest, XGBoost, and CatBoost are at the top; SVM presents good but somewhat lower results, while KNN and Naive Bayes show more limited accuracy.

The ROC-AUC analysis confirms that the observed improvements in the main metrics (such as F1-score, precision, and recall) do not depend on the choice of the decision threshold, but represent a real ability of the models to accurately distinguish crystals with different elastic properties.

5. Conclusions

This paper implemented an ML framework for binary classification of elastic properties in crystalline materials, with a particular focus on addressing the challenge of class imbalance. Using the elastic_tensor_2015 dataset from the Materials Project and implementing a systematic comparison of eight machine learning algorithms and two advanced resampling techniques, several key conclusions can be drawn.

The research shows that proper feature engineering is essential for building predictive models in materials informatics. We used only the basic material characteristics, such as chemical composition and crystal structure, and did not include previously calculated elastic properties in the model. Thus, the model is not learning "paralogous" or direct relationships, but is predicting elastic behaviour in a realistic way. As a result, the performance measures (F1 = 0.697–0.789) accurately reflect the difficulty of predicting elastic properties from basic material characteristics alone.

Second, our comparison showed that ensemble methods outperform other approaches for this classification task. In particular, XGBoost gave the highest performance when used with ADASYN (F1 = 0.789, ROC-AUC = 0.863), followed by Balanced Random Forest and CatBoost. This indicates that tree-based ensemble methods are particularly effective in capturing complex and nonlinear relationships between basic material characteristics and elastic properties of crystals.

Third, the comparison between SMOTE-Tomek and ADASYN showed that both methods improved the handling of class imbalance, but ADASYN gave better performance for ensemble methods. This is because ADASYN focuses on creating synthetic samples for minority examples that are harder to learn, which is particularly helpful for tree-based models. On the other hand, SMOTE-Tomek provided more consistent results across different types of algorithms, making it a reliable choice for general use.

Fourth, the feature importance analysis showed that some specific Magpie features, especially those related to electronic structure and atomic properties, have the greatest impact on the prediction of Poisson's ratio. This shows that machine learning models can provide valuable scientific insights and directly relate their predictions to the physical behaviour of materials, making the model not only accurate but also easy to interpret.

In summary, this study provides a reliable and scientifically sound framework for predicting elastic properties, avoiding errors from data leakage, and providing practically useful results. The findings provide clear guidance for materials scientists, suggesting:

- The use of ensemble methods, especially XGBoost and Balanced Random Forest, as the main choice.
- The use of ADASYN resampling to maximise performance with ensemble methods.
- A comprehensive feature engineering based on composition and structure.
- Realistic performance expectations for real predictive tasks.

Although the presented model has given very good results, there are some limitations that need to be considered. First, the dataset used (elastic_tensor_2015) is not very large and only includes material structures under standard conditions. This may limit the model's ability to be used on more complex materials or under other conditions, such as changes in temperature or pressure.

Secondly, the models are based mainly on data on the composition and structure of the material, without including other factors such as defects or microstructural features that may affect the elastic behaviour.

This study opens up some interesting avenues for further research. One of them is the use of more advanced methods for synthetic data generation, such as Generative Adversarial Network (GAN) models. Another direction is to extend the applications of the validated framework to predict other properties of materials with unbalanced distributions, such as thermodynamic stability or electronic conductivity, moving from binary classification to multiclass tasks or regression with a steep

distribution. Also, the use of advanced interpretation tools, such as SHAP analysis, can extract valuable insights and reveal design rules from high-performance “black box” models, thus linking predictions to scientific meaning.

Acknowledgment:

This work has been supported by the European COST Action EuMINe – European Materials Informatics Network (CA22143), under the COST (European Cooperation in Science and Technology) framework. COST (European Cooperation in Science and Technology) is a funding agency for research and innovation networks; its Actions help connect research initiatives across Europe and foster collaboration. (www.cost.eu)

References:

[1] Azad, Muhammad Muzammil, Sungjun Kim, and Heung Soo Kim. "Autonomous data-driven delamination detection in laminated composites with limited and imbalanced data." Alexandria Engineering Journal 107 (2024): 770-785.

[2] Choudhary, Kamal, Brian DeCost, Chi Chen, Anubhav Jain, Francesca Tavazza, Ryan Cohn, Cheol Woo Park et al. "Recent advances and applications of deep learning methods in materials science." npj Computational Materials 8, no. 1 (2022): 59.

[3] Fang, Zirui, Zixuan Li, Ming Li, Zhenyu Yue, and Ke Li. "Prediction of Protein-DNA Interface Hot Spots Based on Empirical Mode Decomposition and Machine Learning." Genes 15, no. 6 (2024): 676.

[4] Farahnakian, Fahimeh, Javad Sheikh, Luca Zelioli, Dipak Nidhi, Iiro Seppä, Rami Ilo, Paavo Nevalainen, and Jukka Heikkonen. "Addressing imbalanced data for machine learning based mineral prospectivity mapping." Ore Geology Reviews (2024): 106270.

[5] Gao, Chaochao, Xin Min, Minghao Fang, Tianyi Tao, Xiaohong Zheng, Yangai Liu, Xiaowen Wu, and Zhaohui Huang. "Innovative materials science via machine learning." Advanced Functional Materials 32, no. 1 (2022): 2108044.

[6] Jaafreh, R., J. G. Kim, and K. Hamad. Interpretable Machine Learning Analysis of Stress Concentration in Magnesium: An Insight beyond the Black Box of Predictive Modeling. Crystals 2022, 12, 1247. 2022.

[7] Jiang, Jian, Chunhuan Zhang, Lu Ke, Nicole Hayes, Yueying Zhu, Huahai Qiu, Bengong Zhang, Tianshou Zhou, and Guo-Wei Wei. "A review of machine learning methods for imbalanced data challenges in chemistry." Chemical Science (2025).

[8] Mohosheu, Md Salman, and Md Abdullah Al Noman. "Learning from Imbalanced Data." PhD diss., Department of Electrical and Elecrtonics Engineering (EEE), Islamic University of Technology (IUT), Board Bazar, Gazipur-1704, Bangladesh, 2024.

[9] Song, Yimo, Xiangnong Wu, Kexin Wang, and Yiwen Zhang. "Discovering optical performance enhancers in perovskite materials through machine learning-based feature analysis." Optics Express 33, no. 1 (2025): 776-787.

[10] Werner de Vargas, Vitor, Jorge Arthur Schneider Aranda, Ricardo dos Santos Costa, Paulo Ricardo da Silva Pereira, and Jorge Luis Victória Barbosa. "Imbalanced data preprocessing techniques for machine learning: a systematic mapping study." Knowledge and information systems 65, no. 1 (2023): 31-57.

[11] Xu, Pengcheng, Xiaobo Ji, Minjie Li, and Wencong Lu. "Small data machine learning in materials science." npj Computational Materials 9, no. 1 (2023): 42.

[12] Yadav, Pooja, S. C. Sharma, Rajesh Mahadeva, and Shashikant P. Patole. "Exploring hyper-parameters and feature selection for predicting non-communicable chronic disease using stacking classifier." IEEE Access 11 (2023): 80030-80055.

[13] A. Jain et al., "The Materials Project: A materials genome approach to accelerating materials innovation," APL Mater., vol. 1, no. 1, p. 011002, 2013, doi: 10.1063/1.4812323.