Machine Learning Framework for Characterizing Marine Microbial Enzymes with Microplastic Biodegradation Potential

Djadjiti Namla[1,4], Christian Nwankwo Chijioke[3], Mohammad Oves[2*], Naser Ahmed Alkenani[1], and Majed Al-Shaeri[1]

[1]Department of Biological Sciences, Faculty of Sciences, King Abdul Aziz University, Jeddah 21589, Saudi Arabia

[2]Centre of Excellence in Environmental Studies, King Abdulaziz University, Jeddah 21589, Saudi Arabia

[3]National Centre for Artificial Intelligence and Robotics, National Information Technology Development Agency, Abuja, Nigeria

[4]Department of Biotechnology, Faculty of Sciences, Nile University of Nigeria, Abuja, Nigeria

Corresponding author: Mohammed Oves (mmateenuddin@kau.edu.sa)

Abstract

The goal of this study was to develop and evaluate a machine learning framework that can characterize the function of marine microbial enzymes with microplastic (MP) biodegradation potential—based on optimal temperature, pH, and Enzyme Commission number (EC) classification. Data were curated from the three major biological repositories, NCBI, EBI, and DDBJ, in order to target marine bacterial and fungal isolates with recorded enzyme activity and environmental metadata. After extensive curation, cleaning, feature engineering, and handling missing values, a refined dataset comprising 167 samples and seven features was prepared for modeling. Class imbalance was addressed by applying the Synthetic Minority Oversampling Technique (SMOTE), and the resultant dataset was further divided into 80% training and 20% testing subsets. Two ensemble models were assessed—and these include the Gradient Boosting Classifier and Random Forest. Each was evaluated with several performance metrics. Indeed, the Random Forest model outperformed the best-performing model with an accuracy of 0.79 and superior weighted F1-score of 0.72 compared with the general performance of the Gradient Boosting model at an accuracy of 0.70. Despite challenges presented by the underrepresented enzyme classes, the Random Forest yielded good predictive

capabilities for the dominant enzyme functions. This framework provides a foundational step toward predictive bioprospecting were identification of marine microbial enzymes with desirable biochemical properties, such as thermostability or specific catalytic activity, can be achieved. Overall, the study presents a scalable computational approach toward accelerating the discovery and functional characterization of novel marine enzymes for biotechnological and industrial applications.

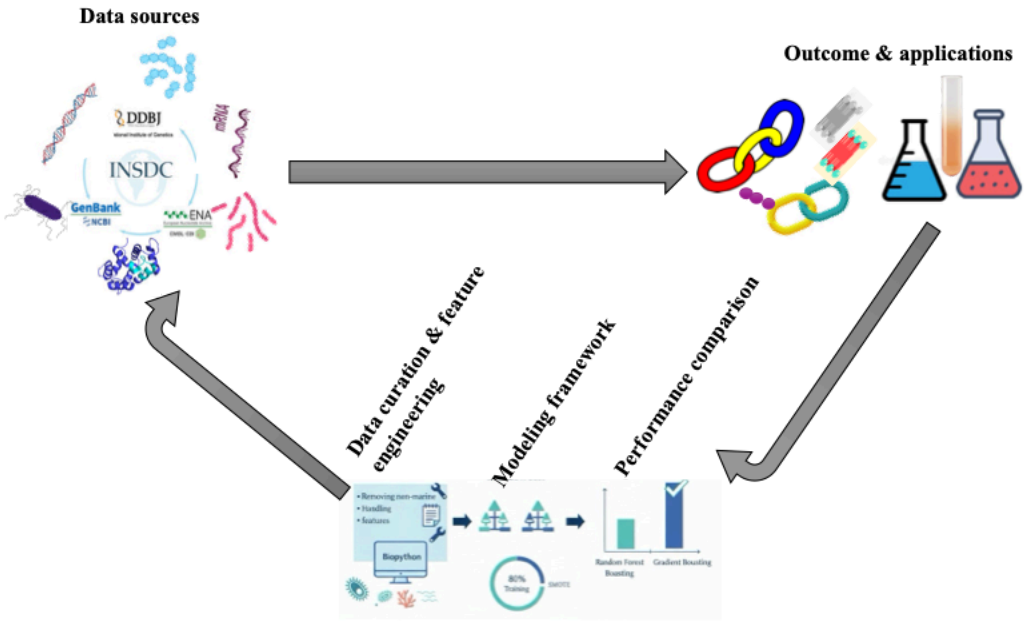
Keywords:

Machine learning, Artificial intelligence, Marine, Bacteria, Fungi, Microalgae, Microbial enzyme.

Highlights:

- The function of microbial enzymes of marine origin possessing potential for microplastic biodegradation, were characterized using a machine learning framework, optimally with respect to temperature and pH, using the Enzyme Commission classification.
- Data were curated representing the three major biological repositories, NCBI, EBI, and DDBJ, targeting marine bacterial, fungal, and microalgal isolates with recorded enzyme activity along with associated environmental metadata.
- The EC number was decomposed into four new features: ec_1, ec_2, ec_3, and ec_4 to enhance the model's efficiency in enzyme function characterization.
- Highly imbalanced distribution of enzyme functions was addressed using SMOTE on the training set to avoid biased learning.
- The Random Forest model outperformed the Gradient Boosting Classifier on the test set, with an accuracy of 0.79 and a weighted F1-score of 0.72.

Graphical Abstract



1. Introduction

The global rise in microplastic pollution has intensified the search for sustainable biodegradation solutions and marine microbial communities have emerged as promising but underexplored reservoirs of plastic-degrading enzymes [1,2,3,4,5]. Recent developments in artificial intelligence (AI) and machine learning (ML) have completely reshaped bioprospecting strategies through the rapid, data-driven discovery of novel enzymatic functions in complex biological and environmental datasets [6,7]. In marine systems, where microbial diversity is vast and genomic information is expanding rapidly [8], AI-driven methods offer an unprecedented opportunity to accelerate the identification and characterization of enzymes with biotechnological and ecological relevance [9].

Traditional bioprospecting methods for marine enzymes rely on culturing, biochemical screening [10,11], and targeted sequencing—processes that are labor-intensive and often limited by the unculturability of many marine microbes

[12,13,14]. AI and ML mitigate these constraints by learning functional patterns from curated datasets, integrating environmental metadata, and generating predictive insights that can guide laboratory validation [15,16]. Through these approaches, researchers can model enzyme–environment relationships, classify functional groups, and identify microbial taxa with high potential for biotechnological exploitation [17,18].

ML has become instrumental in processing heterogeneous biological datasets, from genomic descriptors to physicochemical parameters [19]. In enzyme characterization studies, ML models can infer function from sequence-derived features, optimal activity conditions [20], and ecological context, while also enabling the efficient prioritization of candidate enzymes [21]. The dataset attached in the supplementary information exemplifies the need for AI-assisted preprocessing and feature engineering—including handling missing environmental metadata, standardizing taxonomic nomenclature, and decomposing Enzyme Commission numbers (EC) into interpretable hierarchical features [22,23]. Each of these steps strengthens downstream classification tasks in support of robust prediction of enzyme properties from minimal or noisy inputs [24].

A data-driven prediction pipeline enables the systematic assessment of marine bacterial enzymes involved in microplastic biodegradation [25,26]. Ensemble ML methods, such as Random Forest and Gradient Boosting, are particularly effective for small, heterogeneous biological datasets owing to their robustness to noise and their capability of modeling nonlinear relationships [27,28,29]. Models trained on curated features including environmental parameters ssuch as temperature, pH, salinity and some meta data such as location could reach high predictive performance, with Random Forest imbalance [30,31]. This kind of framework could easily be used for predictions of degradative potential, identification of promising microbial isolates, and enzyme discovery, including the Red Sea for sustainability applications.

The Red Sea is a unique marine system with high salinity, elevated temperatures, and diverse microbial assemblages adapted to such extremes. These features make the region of high interest in the search for robust enzymes tolerant of such harsh industrial and environmental conditions required to ensure the effectiveness of microplastic biodegradation [32,33,34]. In support of targeted bioprospecting for enzymes with functional relevance in plastic mitigation, this study integrates Red Sea microbial genomic and environmental data into an ML-based characterization pipeline. The framework will herein contribute to both fundamental microbial ecology and applied biotechnology, centered on enabling scalable and nature-inspired solutions to global plastic pollution.

2. Materials and Methods

Data used in this study were secondary data curated from three international biological repository data banks known as GenBank. The selected databases include the National Center for Biotechnology Information (NCBI, USA), the European Bioinformatic Institute (EBI) and The DNA Data Bank of Japan (DDBJ).

2.1. Data Acquisition

We targeted marine origin bacterial and fungal entries reporting including (i) taxonomic identification (scientific/binomial name and, where available, strain), (ii) enzyme(s) produced with Enzyme Commission (EC) numbers, (iii) gene(s) encoding the enzyme(s), (iv) optimal temperature and pH for enzyme activity (or growth when activity data were absent but explicitly linked), (v) functional annotation of the enzyme (free text and controlled terms), and (vi) stable accession identifiers (GenBank, EBI and DDBJ). Entries without clear marine provenance or lacking an accession were excluded.

Then queries of interest from the biodabases were executed programmatically either from the GenBank site's interface using the "Advanced search" menu or in Python using Biopython's Entrez and EBI/ENA REST endpoints with Boolean

expressions combining marine context and enzyme terms. Example query templates:

▪ Taxa/environment: "marine"[All Fields] OR seawater OR sediment OR "hydrothermal vent" OR "coral reef".

▪ Enzyme filter: EC: OR "lyase" OR "protease" OR "lipase" OR "amylase" OR "cellulase" OR "chitinase" OR "alginate lyase".

▪ Combined example (NCBI Nucleotide/BioSample): (marine OR seawater OR sediment) AND (EC[All Fields] OR lyase OR protease OR lipase OR amylase OR cellulase OR chitinase).

A well-structured bioinformatic pipeline for developing an explicit dataset of enzyme-producing marine microorganisms was developed. This served to mine the large International Nucleotide Sequence Database Collaboration repositories for sequences with specific marine-derived enzymes and associated environmental data from bacteria, fungi, and microalgae. This automated workflow ensures that the collection is reproducible with a targeted approach. The complete workflow, from data source to metadata extraction is illustrated in the Figure 1 below.

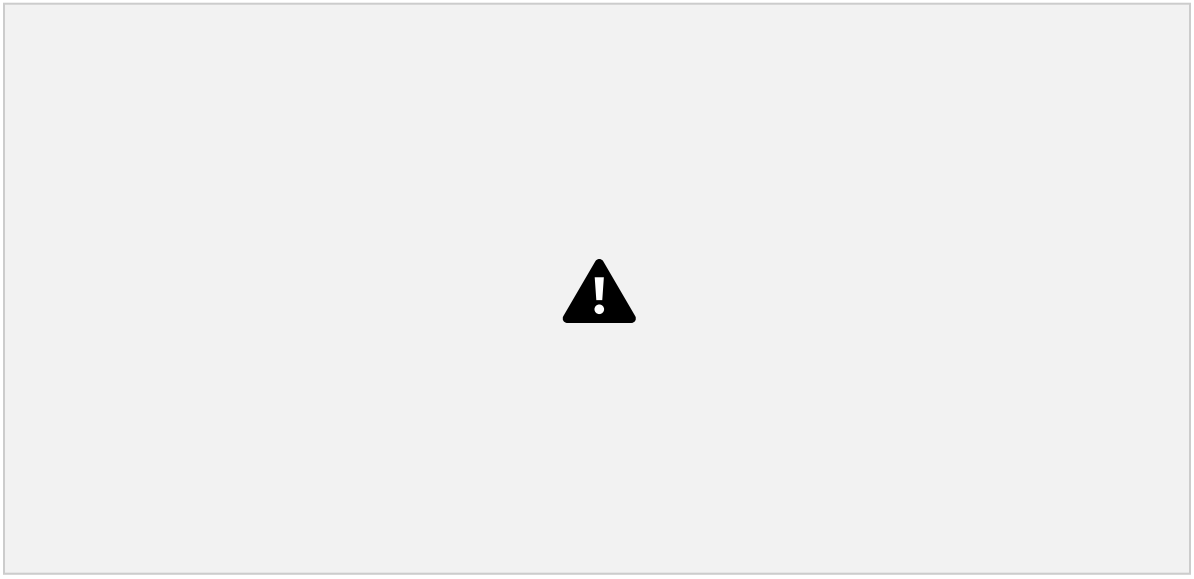


Figure 1. Shematic overview of the bioinformatic workflow. Data were retrieved from the GenBank, ENA, and DDBJ using Biopython queries. The queries were designed to filter for sequences from enzyme-producing marine microbes (including bacteria, fungi and microalgae) isolated from seawater or sea-sediment. Then the retrieved data were processed to extract and organize key metadata fields including name of microorganism, functional classification using the EC numbers, gene encoding, name of enzyme, and associated environmental parameters such as the temperature, geolocalization, and pH.

Some of the selected entries were having genome or gene sequences; and for these, their corresponding sequence records (FASTA/GenBank flat files) and feature tables were downloaded to enable sequence-derived features. This approach yielded a curated dataset linking source organisms, enzyme classifications, and key physicochemical parameters.

However, any isolate with metadata having at least one enzyme with EC number falling within predefined industrial or biotechnology categories (such as hydrolases on polysaccharides, lipases/esterases, proteases, oxidoreductases—with bioremediation relevance), was termed as “Positive” for beneficial enzyme-products. Whereas, if isolates without such evidence were labeled “Negative” Non-reported” and retained for semi-supervised analysis.

2.2. Exploratory Data Analysis

An exploratory data analysis (EDA) was conducted on the dataset to reveal feature distributions and data quality issues, following the iterative workflow depicted in the Figure 3 below. In the course of this, we identified two missing entries in the “Source of Bacteria” column. These rows were dropped as the source is critical identifier that cannot be easily determined by the mean or mode of the distribution.

Ten entries marked “N/R” “Not-reported” in the “pH for Maximum Functionality” column were replaced with NaN and subsequently imputed using the mean pH value of the dataset to retain the data points for analysis.

The “Temperature for Maximum Functionality” column was inconsistent and contained non-numeric characters like “ºC” and entries like “N/R”. These were stripped and converted to a nullable integer type. Eight entries from the bacterium *Zobellia galactanivora* were missing temperature data. Instead of using the global mean, we take a more biologically relevant approach. For the imputation, the mean optimal temperature was input for others enzyme produced by *Zobellia galactanivorans,* in order to preserve the ecological context

Finally, data discripencies on the “Source of microbes” were spotted out and addressed accordingly. For instance, “Vibrio species” was changed to “Vibro sp”, while Zobellia galactanivorans” was standardized to “*Zobellia galactanivora*”, enabling consistent grouping.

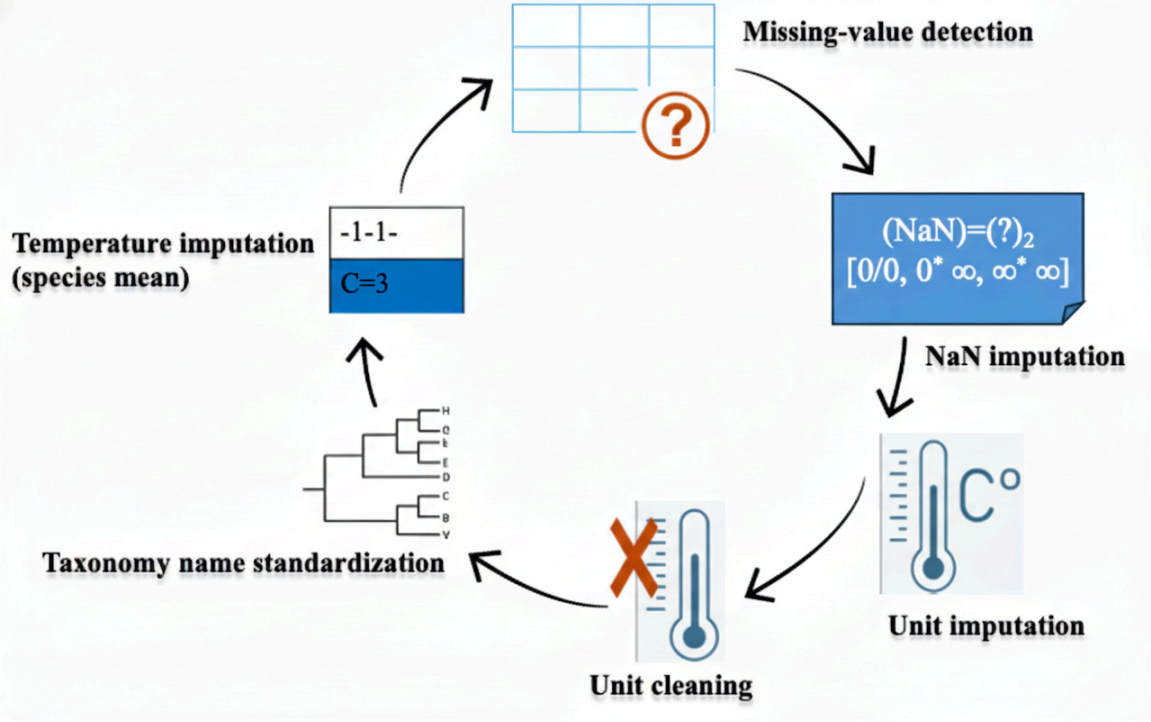


Figure 2. Iterative workflow for exploratory data analysis, outlining the key steps taken to clean, standardize, and prepare the dataset for model development.

2.3. Feature Engineering

To enhance the efficiency of the model in characterizing enzyme function, the EC Number was decomposed into its constituent parts. The EC number is a hierarchical code

where EC_1 represents the main class of reaction, EC_2 indicates the subclass, EC_3 signifies the sub-subclass, and EC_4 is the serial number for the individual enzyme.

This single column EC Number was split into four new features. Specifically, it was split into ec_1, ec_2, ec_3, and ec_4. These new features were converted to numeric types for machine learning compatibility. Twelve rows of missing values in ec_4 were noted and retained for the model to handle. This was done by assigning zero to all the missing ec_4 values.

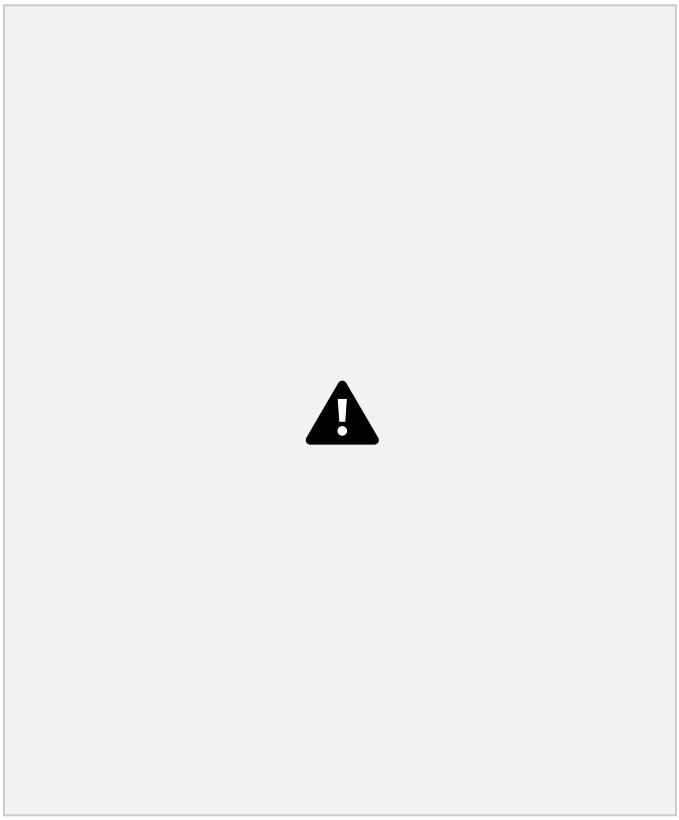


Figure 3. Features engineering process workflow

2.4. Final Dataset for Modeling

Following the cleaning and feature engineering steps, we refined the dataset for modeling. Columns not directly useful for a generalizable model, such as S/N, Enzyme Name, Source of Bacterium or fungus, Gene Encoding, and Organism's Accession Number, were dropped. The final feature set used for the model training consisted of: Temperature for Maximum Functionality, pH for Maximum Functionality, Function,

ec_1, ec_2, ec_3, and ec_4. Except Function that was categorical, the features were all numerical. The Function column is encoded using LabelEncoder from the scikit-learn library. Although Function is a nominal categorical variable that would typically benefit from one-hot encoding to avoid implying ordinal relationships, we use LabelEncoder here because tree-based ensemble models such as Random Forests and Gradient Boosting are not affected by the arbitrary numerical ordering created by label encoding, as these models make splits based on feature values rather than their numerical magnitude. Additionally, label encoding maintains a single column, whereas one-hot encoding would create multiple binary columns, which could significantly increase the feature space depending on the number of unique functions.

This data curation yielded 167 samples and 7 features, providing a good foundation for training our machine learning models to predict enzyme function based on their optimal conditions and EC Number classification.

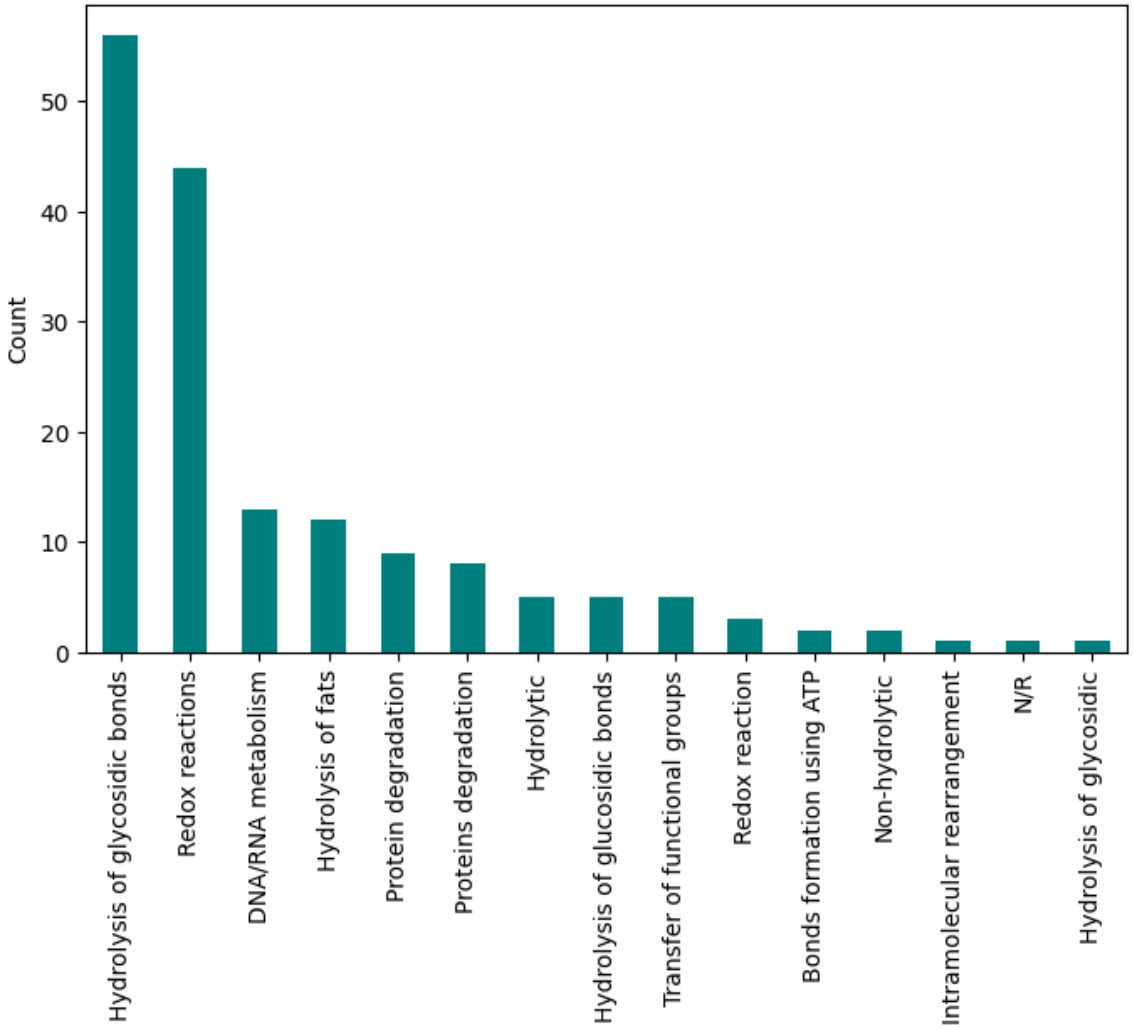


Figure 4. Visual representation of the distribution of enzyme functions.

3. Data Preprocessing and Model Development

This section describes data cleaning and feature engineering, strategies used for addressing class imbalance, the model training pipeline, and the strategy for performance evaluation.

3.1. Data Preprocessing

The dataset was split into a training set and a testing set. We used the standard and conventional split of 80% for training and 20% for testing. This ensured that the models were trained on one subset of data and their performance were evaluated on an unseen subset, providing an unbiased estimate of their generalizability. The classification task relies on a robust machine learning framework that is planned in such a way as to

maximize predictive performance and interpretability. Enzyme functions that appeared only once in the dataset were removed. This step was necessary to enable stratified sampling, ensuring that each enzyme function was represented in both the training and testing sets. After imputation and encoding, the class imbalance issue was dealt with before training the model to avoid biased learning.

3.2. Handling Class Imbalance

In our dataset, the distribution of the enzyme function was highly imbalanced as shown in Figure 1. Class imbalance can lead to biased model predictions, as machine learning algorithms tend to favor majority classes [35]. To address this issue, SMOTE was then applied to the training set only to create synthetic minority-class examples, using $k = 5$ to avoid information leakage into the test set [36]. With this approach, we generated synthetic samples for minority classes by interpolating between existing instances, thereby increasing their representation without duplicating samples. This approach improved the performance of our models across classes. Because preprocessing changes class representation and feature distributions, we subsequently applied a targeted resampling strategy.

3.3. Model Selecction and Training

Given that the task is multi-task classification problem, a suite of models with different inductive biases was selected for evaluation. Considering that our dataset small, previously highly imbalanced, and contained noise, we employed the Gradient Boosting Classifier and Random Forest, both of which are ensemble models. These models were selected because their ensemble structure mitigates the high variance often associated with single decision trees, while maintaining computational efficiency. Ensemble models combine multiple individual learners to improve predictive performance and reduce the risk of overfitting [37,38,39,40]. Figure 1 illustrates the conceptual architecture that describes the flow from data acquisition to performance validation.

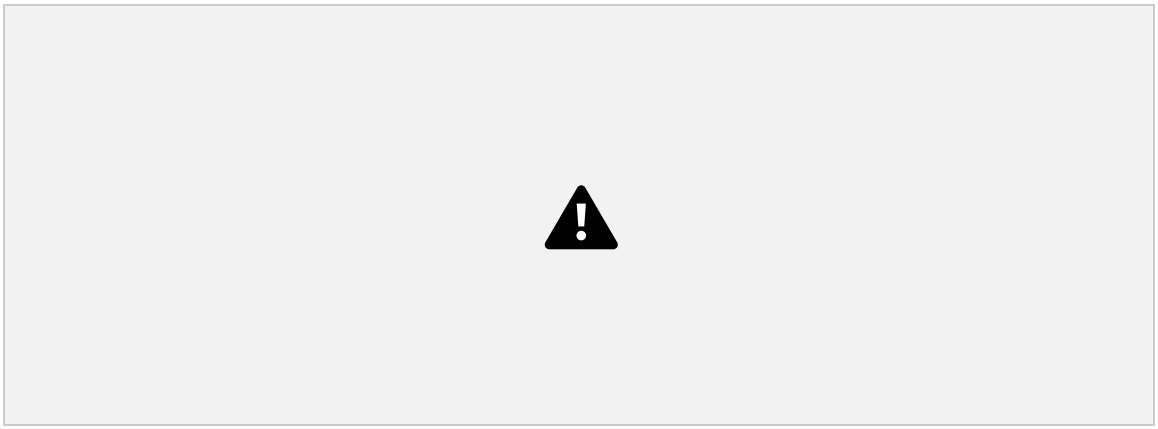


Figure 5. Conceptual framework for ensemble learning classification and evaluation using Random Forest and Gradient Boosting classifiers.

3.4. Evaluation Metrics and Validation Strategy

To effectively assess model performance, multiple evaluation metrics were employed. This is because accuracy alone can be misleading for multi-class problems, especially with potential class imbalance [41].

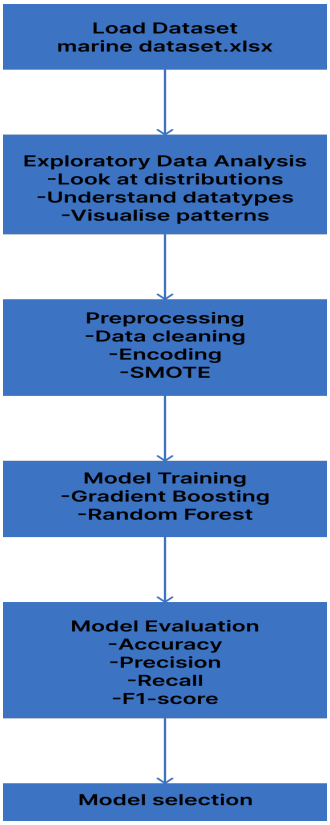


Figure 6. Visual representation of the overall machine learning pipeline, from loading the dataset to model evaluation and selection.

4. Results and Discussion

4.1. Model Performance Comparison

Table 1. Summary of the performance from the Gradient Boosting Classifier model on the test set.

Accuracy: 0.696969696969697

Gradient Boosting Classifier Classification Report:

	precision	recall	f1-score	support
1	0.00	0.00	0.00	3
2	1.00	1.00	1.00	2
3	0.00	0.00	0.00	1
5	0.73	0.73	0.73	11
6	0.33	1.00	0.50	1
9	0.00	0.00	0.00	0
10	1.00	1.00	1.00	2
11	1.00	1.00	1.00	2
12	0.00	0.00	0.00	1
13	0.73	0.89	0.80	9
14	0.00	0.00	0.00	1
accuracy			0.70	33
macro avg	0.44	0.51	0.46	33
weighted avg	0.63	0.70	0.66	33

The model achieved an overall accuracy of 0.70, indicating that approximately 70% of enzyme functions were correctly classified. However, performance varied considerably across classes. As show in the classification report in Table 1, the model achieved strong predictive performance for the more represented enzyme functions with F1-scores of 0.73 and 0.80, respectively. Conversely, classes with very few samples were not correctly predicted. The macro-averaged F1-score (0.46) further reflects this imbalance.

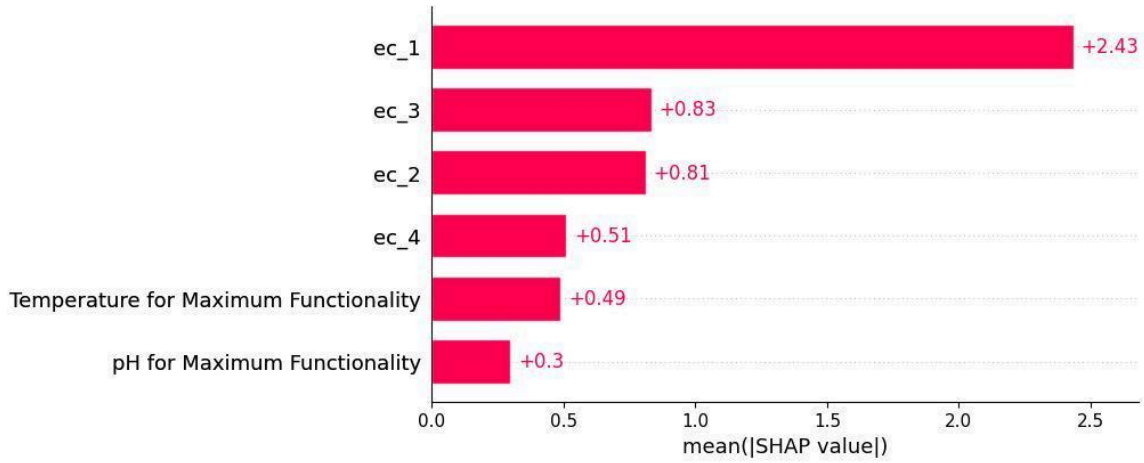


Figure 7. Feature analysis from gradient boosting model, showing ec_1 as the most influential factor in the model's prediction by far.

Table 2. Summary of the performance from the Random Forest model on the test set.

Accuracy: 0.7878787878787878

Classification Report:

	precision	recall	f1-score	support
1	0.00	0.00	0.00	3
2	1.00	1.00	1.00	2
3	0.00	0.00	0.00	1
5	0.79	1.00	0.88	11
6	0.33	1.00	0.50	1
10	1.00	1.00	1.00	2
11	1.00	1.00	1.00	2
12	0.00	0.00	0.00	1
13	0.80	0.89	0.84	9
14	0.00	0.00	0.00	1
accuracy			0.79	33
macro avg	0.49	0.59	0.52	33
weighted avg	0.67	0.79	0.72	33

The Random Forest model achieved an overall accuracy of 0.79, which represents an improvement over the Gradient Boosting Classifier (0.70). As shown in the classification report in Table 2, the model demonstrated high predictive performance for the dominant enzyme functions, with F1-scores of 0.88 and 0.84, respectively. However, enzyme functions with very few samples were not correctly predicted, resulting in undefined precision values for those classes.

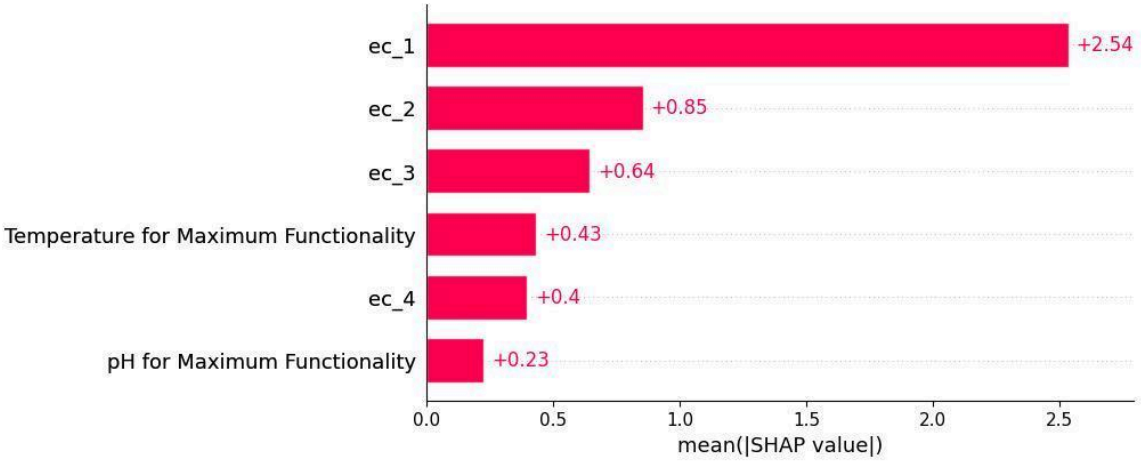


Figure 8. Feature analysis from random forest model, showing ec_1 as the most influential feature in the model's prediction by far.

4.2. Interpretation of Findings and Limitations

The weighted average (0.66) suggests that the model performs adequately for the dominant classes. These results indicate that data scarcity and class imbalance affected the model's ability to generalize across all enzyme functions.

The macro-averaged F1-score (0.72) indicates that the model performed reliably for the majority classes. Overall, the Random Forest outperformed the Gradient Boosting model, showing better generalization despite the dataset's small size and residual imbalance.

4.3. Reprodducibility and Code Availability

All codes, processed datasets and model checkpoints involved in the present study are archived at GitHub via this link [https://github.com/16262chris/Enzyme-producing-marine-bacteria.git], to enable reprodduction of our analyses.

5. Conclusion

This study successfully developed a machine learning framework capable of characterizing the function of marine bacterial enzymes based on their optimal conditions and EC number. The Random Forest Model proved to be a powerful tool for this task. This data-driven approach can help in bioprospecting. For example, if one is searching for a thermostable protease, the model can help identify EC numbers and source organisms likely to produce enzymes with that combined set of properties.

This work lays the foundation for a predictive tool that can accelerate the discovery and characterization of novel enzymes from the vast and largely untapped resource of marine biodiversity.

Reference

1. Zhao, S., Liu, R., Lv, S., et al. (2024). Polystyrene-degrading bacteria in the gut microbiome of marine benthic polychaetes support enhanced digestion of plastic fragments. Communications Earth & Environment, 5, 162. https://doi.org/10.1038/s43247-024-01318-6
2. Lebepe, J., Buthelezi, N. M. D., & Manganyi, M. C. (2025). Occurrence and Control of Microplastics and Emerging Technological Solutions for Their Removal in Freshwaters: A Comprehensive Review. Microplastics, 4(4), 70. https://doi.org/10.3390/microplastics4040070
3. Bogdanović, T., Listeš, I., Gjerde, J., Petričević, S., Jažo, Z., Listeš, E., Pleadin, J., Sokolić, D., Jadrešin, I., & di Giacinto, F. (2025). Microplastic Polymer Mass Fractions in Marine Bivalves: From Isolation to Hazard Risk. Journal of Xenobiotics, 15(6), 186. https://doi.org/10.3390/jox15060186
4. Tan, T., Liu, A., Yang, Y., Yu, R., Lin, N., Ke, Q., & Wang, Q. (2025). Microplastic Pollution in Typical Subtropical Rivers in Eastern China: A Case Study of the Feiyun River Basin. Water, 17(21), 3170. https://doi.org/10.3390/w17213170
5. Nwakanma, S. C., Aransiola, S. A., Adejokun, E. K., Okwute, O. L., Aljerf, L., & Maddela, N. R. (2025). Microbial relationship of marine microplastics. In S. A. Aransiola, B. R. Babaniyi, N. R. Maddela, & O. S. Bello (Eds.), White pollution:

Biodiversity and hazards in marine plastisphere (pp. xx–xx). Environmental contamination remediation and management. Springer. https://doi.org/10.1007/978-3-031-95547-1_7

6. Sharma, D., & Singh, K. (2025). AI-enhanced bioprocess technologies: machine learning implementations from upstream to downstream operations. World journal of microbiology & biotechnology, 41(8), 278. https://doi.org/10.1007/s11274-025-04494-5
7. Mirakhori, F., & Niazi, S. K. (2025). Harnessing the AI/ML in Drug and Biological Products Discovery and Development: The Regulatory Perspective. Pharmaceuticals, 18(1), 47. https://doi.org/10.3390/ph18010047
8. Chen, J., Jia, Y., Sun, Y., Liu, K., Zhou, C., Liu, C., Li, D., Liu, G., Zhang, C., Yang, T., Huang, L., Zhuang, Y., Wang, D., Xu, D., Zhong, Q., Guo, Y., Li, A., Seim, I., Jiang, L., Wang, L., … Fan, G. (2024). Global marine microbial diversity and its potential in bioprospecting. Nature, 633(8029), 371–379. https://doi.org/10.1038/s41586-024-07891-2
9. Yu, T., Chae, M., Wang, Z., Ryu, G., Kim, G. B., & Lee, S. Y. (2025). Microbial Technologies Enhanced by Artificial Intelligence for Healthcare Applications. Microbial biotechnology, 18(3), e70131. https://doi.org/10.1111/1751-7915.70131
10. Lv, S., Wang, Q., Li, Y., Gu, L., Hu, R., Chen, Z., & Shao, Z. (2024). Biodegradation of polystyrene (PS) and polypropylene (PP) by deep-sea psychrophilic bacteria of Pseudoalteromonas in accompany with simultaneous release of microplastics and nanoplastics. The Science of the total environment, 948, 174857. https://doi.org/10.1016/j.scitotenv.2024.174857
11. Zhivkoplias, E., Jouffray, J. B., Dunshirn, P., et al. (2024). Growing prominence of deep-sea life in marine bioprospecting. Nature Sustainability, 7, 1027–1037. https://doi.org/10.1038/s41893-024-01392-w
12. Rodrigues, C. J. C., & de Carvalho, C. C. C. R. (2022). Marine Bioprospecting, Biocatalysis and Process Development. Microorganisms, 10(10), 1965. https://doi.org/10.3390/microorganisms10101965
13. Tatta, E. R., Imchen, M., Moopantakath, J., & Kumavath, R. (2022). Bioprospecting of microbial enzymes: current trends in industry and healthcare.

Applied microbiology and biotechnology, 106(5-6), 1813–1835. https://doi.org/10.1007/s00253-022-11859-5

14. Soccol, C. R., Colonia, B. S. O., de Melo Pereira, G. V., Mamani, L. D. G., Karp, S. G., Thomaz Soccol, V., Penha, R. O., Dalmas Neto, C. J., & César de Carvalho, J. (2022). Bioprospecting lipid-producing microorganisms: From metagenomic-assisted isolation techniques to industrial application and innovations. Bioresource technology, 346, 126455. https://doi.org/10.1016/j.biortech.2021.126455
15. Jiang, R., Shang, L., Wang, R., Wang, D., & Wei, N. (2023). Machine learning based prediction of enzymatic degradation of plastics using encoded protein sequence and effective feature representation. Environmental Science & Technology Letters, 10(7), 557–564. https://doi.org/10.1021/acs.estlett.3c00293
16. Marti, T. D., Schweizer, D., Yu, Y., Schärer, M. R., Probst, S. I., & Robinson, S. L. (2024). Machine learning reveals signatures of promiscuous microbial amidases for micropollutant biotransformations. ACS Environ. Au, 5(1), 114–127. https://doi.org/10.1021/acsenvironau.4c00066
17. Jiang, R., Yue, Z., Shang, L., Wang, D., & Wei, N. (2024). PEZy-miner: An artificial intelligence driven approach for the discovery of plastic-degrading enzyme candidates. Metabolic engineering communications, 19, e00248. https://doi.org/10.1016/j.mec.2024.e00248
18. Wang, Z., Xie, D., Wu, D., Luo, X., Wang, S., Li, Y., Yang, Y., Li, W., & Zheng, L. (2025). Robust enzyme discovery and engineering with deep learning using CataPro. Nature communications, 16(1), 2736. https://doi.org/10.1038/s41467-025-58038-4
19. Han, Y., Zhang, H., Zeng, Z., Liu, Z., Lu, D., & Liu, Z. (2024). Descriptor-augmented machine learning for enzyme-chemical interaction predictions. Synthetic and systems biotechnology, 9(2), 259–268. https://doi.org/10.1016/j.synbio.2024.02.006
20. Norton-Baker, B., Komp, E., Gado, J. E., Denton, M. C. R., Mathews, I. I., Murphy, N. P., Erickson, E., Storment, O. O., Sarangi, R., Gauthier, N. P., McGeehan, J. E., & Beckham, G. T. (2025). Machine Learning-Guided

Identification of PET Hydrolases from Natural Diversity. ACS catalysis, 15(18), 16070–16083. https://doi.org/10.1021/acscatal.5c03460

21. Zaretckii, M., Buslaev, P., Kozlovskii, I., Morozov, A., & Popov, P. (2024). Approaching Optimal pH Enzyme Prediction with Large Language Models. ACS synthetic biology, 13(9), 3013–3021. https://doi.org/10.1021/acssynbio.4c00465
22. Koukaras, P., & Tjortjis, C. (2025). Data Preprocessing and Feature Engineering for Data Mining: Techniques, Tools, and Best Practices. AI, 6(10), 257. https://doi.org/10.3390/ai6100257
23. van der Weg, K., Merdivan, E., Piraud, M., & Gohlke, H. (2025). TopEC: prediction of Enzyme Commission classes by 3D graph neural networks and localized 3D protein descriptor. Nature communications, 16(1), 2737. https://doi.org/10.1038/s41467-025-57324-5
24. Ratna, H. V. K., Jeyaraman, M., Jeyaraman, N., Nallakumarasamy, A., Sharma, S., Khanna, M., & Gupta, A. (2023). Machine learning and deep neural network-based learning in osteoarthritis knee. World journal of methodology, 13(5), 419–425. https://doi.org/10.5662/wjm.v13.i5.419
25. Li, S., & Zhang, W. (2025). Computational identification of plastic-degrading enzymes in ocean microbiomes. Scientific Reports, 15, 15332. https://doi.org/10.1038/s41598-025-99275-3
26. Ruginescu, R., & Purcarea, C. (2024). Plastic-Degrading Enzymes from Marine Microorganisms and Their Potential Value in Recycling Technologies. Marine Drugs, 22(10), 441. https://doi.org/10.3390/md22100441
27. Kyriazos, T., & Poga, M. (2024). Application of Machine Learning Models in Social Sciences: Managing Nonlinear Relationships. Encyclopedia, 4(4), 1790-1805. https://doi.org/10.3390/encyclopedia4040118
28. Alam, M. N. U., Basava, K., Chitransh, A., Fattah, H. M. A., Garcia-Verdugo, H. D., Lo, S. H., Lohchab, T., Martinet, K. M., Román-Palacios, C., Salazar, J. C., & Van Boxel, D. (2025). Machine learning in biological research: key algorithms, applications, and future directions. BMC biology, 23(1), 324. https://doi.org/10.1186/s12915-025-02424-3

29. Azedou, A., Amine, A., Kisekka, I., & Lahssini, S. (2025). Genetic algorithm optimization of ensemble learning approach for improved land cover and land use mapping: Application to Talassemtane National Park. Ecological Indicators, 177, 113776. https://doi.org/10.1016/j.ecolind.2025.113776
30. Hassan, R., Behtouei, Z., & Baghban, A. (2025). Advanced machine learning for precise prediction of biochar's heavy metal sorption efficiency. Journal of Hazardous Materials Advances, 18, 100739. https://doi.org/10.1016/j.hazadv.2025.100739
31. Alnemari, A. M., Elmessery, W. M., Qazaq, A. S., Moustapha, M. E., Rakhimgaliyeva, S., Abuhussein, M. F. A., Alhag, S. K., Al-Shuraym, L. A., Moghanm, F. S., Szűcs, P., Eid, M. H., & Elwakeel, A. E. (2025). Developing highly accurate machine learning models for optimizing water quality management decisions in tilapia aquaculture. Scientific Reports, 15, 35600. https://doi.org/10.1038/s41598-025-16939-w
32. Renn, D., Shepard, L., Vancea, A., Karan, R., Arold, S. T., & Rueping, M. (2021). Novel Enzymes From the Red Sea Brine Pools: Current State and Potential. Frontiers in microbiology, 12, 732856. https://doi.org/10.3389/fmicb.2021.732856
33. Delgadillo-Ordoñez, N., Raimundo, I., Barno, A. R., Osman, E. O., Villela, H., Bennett-Smith, M., Voolstra, C. R., Benzoni, F., & Peixoto, R. S. (2022). Red Sea Atlas of Coral-Associated Bacteria Highlights Common Microbiome Members and Their Distribution across Environmental Gradients—A Systematic Review. Microorganisms, 10(12), 2340. https://doi.org/10.3390/microorganisms10122340
34. Altalhi, S., Schultz, J., Jamil, T., Diercks, I., Sharma, S., Follmann, J., Alam, I., Raman, K., Augustin, N., van der Zwan, F. M., & Rosado, A. S. (2025). Decoding microbial diversity, biogeochemical functions, and interaction potentials in red sea hydrothermal vents. Environmental microbiome, 20(1), 118. https://doi.org/10.1186/s40793-025-00784-5
35. Albattah, W., & Khan, R. U. (2025). Impact of imbalanced features on large datasets. Frontiers in big data, 8, 1455442. https://doi.org/10.3389/fdata.2025.1455442

36. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953
37. Khan, A. A., Chaudhari, O., & Chandra, R. (2024). A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation. Expert Systems with Applications, 244, 122778. https://doi.org/10.1016/j.eswa.2023.122778
38. Thelagathoti, R. K., Chandel, D. S., Tom, W. A., Jiang, C., Krzyzanowski, G., Olou, A., & Fernando, M. R. (2025). Machine Learning-Based Ensemble Feature Selection and Nested Cross-Validation for miRNA Biomarker Discovery in Usher Syndrome. Bioengineering, 12(5), 497. https://doi.org/10.3390/bioengineering12050497
39. Kassem, M. A. (2025). Harnessing Artificial Intelligence and Machine Learning for Identifying Quantitative Trait Loci (QTL) Associated with Seed Quality Traits in Crops. Plants, 14(11), 1727. https://doi.org/10.3390/plants14111727
40. Jang, Y. J., Yun, D., Shin, W., Goo, C., Song, C. M., Han, K., Kim, S., Kim, D.-S., Lee, S., & Oh, Y. (2025). Integrative Machine Learning Approaches for Identifying Loci Associated with Anthracnose Resistance in Strawberry. Plants, 14(18), 2889. https://doi.org/10.3390/plants14182889
41. Ghanem, M., Ghaith, A. K., El-Hajj, V. G., Bhandarkar, A., de Giorgio, A., Elmi-Terander, A., & Bydon, M. (2023). Limitations in Evaluating Machine Learning Models for Imbalanced Binary Outcome Classification in Spine Surgery: A Systematic Review. Brain sciences, 13(12), 1723. https://doi.org/10.3390/brainsci13121723