

## Deep Learning and Bioengineered Feature Engineering for Automated Taxonomic Identification of Mediterranean and Atlantic Demospongiae

Fatimah Mohammed Albudwi Alghuddi<sup>1,\*</sup>  

<sup>1</sup>Marine Biology, Nalut University, Faculty of Education-Jadu, Libya

### ARTICLE HISTORY

Received 02 March 2026  
Revised 23 March 2026  
Accepted 29 March 2026  
Online 02 April 2026

### KEYWORDS

Demospongiae taxonomy;  
Imbalanced classification;  
Bioengineered features;  
Deep learning;  
Marine biodiversity;  
hierarchical modeling,  
UCI sponge dataset.

### ABSTRACT

Given morphological convergence, cryptic speciation, and severe class imbalance in regional collection records, accurate taxonomic identification of marine Demospongiae remains a major challenge in biodiversity monitoring. This study presents a novel computational framework that integrates imbalanced multi-class deep learning with biologically informed feature engineering to automate the hierarchical classification of 503 Demospongiae specimens collected from Mediterranean and Atlantic habitats. We introduce a biologically grounded feature construction pipeline that encodes morphological, ecological, and evolutionary priors derived from sponge ontology structures, generating high-dimensional representations compatible with deep neural architectures. To address strong distributional skew across 7 orders, 42 families, 114 genera, and 230 species, we implement a hybrid imbalance mitigation strategy combining topology-aware synthetic minority oversampling, adaptive class-weighted sampling, and hierarchical focal loss. The proposed architecture further employs a multi-task graph convolutional network to jointly learn taxonomic relationships while preserving hierarchical constraints. Experimental evaluation demonstrates substantial improvements over conventional machine learning baselines, achieving macro-averaged F1-scores of 0.89 at the order level and 0.76 at the species level, with notable gains in recall for underrepresented Atlantic taxa. Ablation analyses further indicate that the incorporation of bioengineered features significantly enhances model generalization.

## التعلم العميق وهندسة الميزات الحيوية لتحديد التصنيف الآلي للإسفنجيات المتوسطية والأطلسية

فاطمة محمد البدوي<sup>1,\*</sup>

المخلص	الكلمات المفتاحية
نظراً للتقارب المورفولوجي، والتنوع الخفي، وعدم التوازن الحاد في تصنيف الإسفنجيات البحرية في سجلات المجموعات الإقليمية، لا يزال التحديد التصنيفي الدقيق لهذه الإسفنجيات يمثل تحدياً كبيراً في رصد التنوع البيولوجي. تقدم هذه الدراسة إطاراً حسابياً جديداً يدمج التعلم العميق متعدد الفئات غير المتوازن مع هندسة الميزات المستندة إلى أسس بيولوجية لأتمتة التصنيف الهرمي لـ 503 عينة من الإسفنجيات البحرية جُمعت من بيئات البحر الأبيض المتوسط والمحيط الأطلسي. نقدم مساراً لبناء الميزات قائماً على أسس بيولوجية، يشفر المعلومات المورفولوجية والبيئية والتطورية المستمدة من هياكل علم الإسفنج، مما يُنتج تمثيلات عالية الأبعاد متوافقة مع بني الشبكات العصبية العميقة. وللمعالجة الانحراف التوزيعي الكبير عبر 7 رتب، و42 عائلة، و114 جنساً، و230 نوعاً، نطبق استراتيجية هجينة للتخفيف من عدم التوازن تجمع بين أخذ عينات زائدة اصطناعية للأقلية مع مراعاة الطوبولوجيا، وأخذ عينات تكيفية موزونة حسب الفئة، وفقدان التركيز الهرمي. تستخدم البنية المقترحة شبكة عصبية تلافيفية بيانية متعددة المهام لتعلم العلاقات التصنيفية مع الحفاظ على القيود الهرمية. تُظهر التقييمات التجريبية تحسينات كبيرة مقارنةً بنماذج التعلم الآلي التقليدية، حيث حققت متوسطات F1 الكلية 0.89 على مستوى الرتبة و0.76 على مستوى النوع، مع مكاسب ملحوظة في استرجاع البيانات لأنواع الأطلسية الأقل تمثيلاً. وتشير تحليلات الاستبعاد أيضاً إلى أن دمج الميزات المهندسة حيويًا يُحسن بشكل كبير من تعميم النموذج.	تصنيف Demospongiae التصنيف غير المتوازن ميزات الهندسة الحيوية التعلم العميق التنوع البيولوجي البحري النمذجة الهرمية مجموعة بيانات الإسفنج UCI

### Introduction

The phylum Porifera, which includes marine sponges, is a fundamental part of benthic ecosystems, helping with biogeochemical processes, habitat structure, and nutrient cycles [1]. The class Demospongiae, which includes more than 85% of known sponge species, is the most diversified group in this phylum [2]. Expert morphological evaluation of spicule architecture, skeletal organization, and soft-tissue properties is the basis for traditional taxonomic identification

of Demospongiae. This approach is laborious, subjective, and increasingly limited by a lack of taxonomic knowledge [3]. Although molecular barcoding has improved traditional methods, it is still constrained by gaps in reference databases and practical difficulties in remote sampling situations. [4]. There are potential opportunities to automate taxonomic procedures with machine learning. The hierarchical nature of biological classification systems and the extreme class

\*Corresponding author

[https://doi.org/10.63318/waujpasv4i1\\_33](https://doi.org/10.63318/waujpasv4i1_33)

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).



imbalance present in regional biodiversity datasets, where common Mediterranean taxa greatly outnumber rare Atlantic counterparts, are two crucial aspects that are often ignored by current computational approaches. These difficulties are best shown by the UCI Demospongiae dataset [5], which contains 503 specimens from 230 species, 451 of which are from the Mediterranean basin, and only 52 of which are from collections in the Atlantic. Conventional classifiers are predisposed to majority-class bias because to this imbalance, compromising identification accuracy for ecologically significant but underrepresented lineages.

Three integrated contributions are used in this study to close these gaps:

In order to translate sponge ontology structures (orders → families → genera → species) into machine-readable representations that incorporate morphological constraints, biogeographic priors, and phylogenetic closeness metrics, this study established a biologically grounded feature building methodology.

This study created a multi-task graph convolutional network with hierarchical focal loss that prioritizes minority classes by dynamically reweighting gradients and explicitly modeling taxonomic relationships [41-44]. In order to ensure that model performance represents actual deployment situations across Mediterranean-Atlantic gradients, this research has employed a stratified cross-validation technique that maintains biogeographic dispersion patterns. This work advances automated taxonomic identification while offering a transportable framework for unbalanced classification in biodiversity informatics by connecting marine biological expertise with cutting-edge deep learning techniques.

**Table 1:** Demospongiae Dataset Characteristics

Characteristic	Description
<b>Dataset Name</b>	Demospongiae
<b>Source</b>	UCI Machine Learning Repository [5]
<b>Dataset URL</b>	<a href="https://archive.ics.uci.edu/dataset/190/demospongiae">https://archive.ics.uci.edu/dataset/190/demospongiae</a>
<b>Domain</b>	Marine Biology / Taxonomy
<b>Task Type</b>	Hierarchical Multi-class Classification
<b>Total Specimens</b>	503
<b>Format</b>	NOOS (Lisp-like feature-term representation) / Prolog clauses

**Table 2:** Taxonomic Distribution

Taxonomic Rank	Number of Classes	Samples per Class Range
<b>Order</b>	7	42 - 117 specimens
<b>Family</b>	42	1 - 43 specimens
<b>Genus</b>	114	1 - 34 specimens
<b>Species</b>	230	1 - 15 specimens

**Table 3:** Geographic Distribution

Region	Number of Specimens	Percentage
<b>Mediterranean Sea</b>	451	89.70%
<b>Atlantic Ocean</b>	52	10.30%

## Related Work

### Taxonomy of Marine Invertebrates

Rule-based expert systems encoding morphological keys were used in early sponge classification studies [6]. Support vector machines and random forests on manually created morphometric features were used in later machine learning

applications [7], but these methods had trouble with hierarchical relationships and feature scalability. With little focus on poriferan systems, where diagnostic traits are frequently non-visual or necessitate microscopic inspection, recent deep learning applications in marine taxonomy have concentrated on image-based identification of fish [8] or coral [9].

**Table 4:** Feature Characteristics

Feature Attribute	Details
<b>Total Attributes</b>	42 discrete morphological and ecological features
<b>Feature Types</b>	Integer, Categorical
<b>Key Features</b>	Spicule type (monaxon, tetraaxon, polyaxon) Growth form Skeletal organization Habitat depth Body size (mm) Color code Surface texture Oscule diameter Choanocyte chamber type Fiber presence
<b>Missing Values</b>	< 3% of entries
<b>Imputation Method</b>	k-nearest neighbors within taxonomic families

**Table 5:** Class Imbalance Metrics

Metric	Value
<b>Imbalance Ratio (Order)</b>	~2.8:1 (117 max : 42 min)
<b>Imbalance Ratio (Species)</b>	15:1 (15 max : 1 min)
<b>Geographic Imbalance</b>	8.7:1 (Mediterranean : Atlantic)
<b>Minority Classes</b>	114 genera with <10 specimens Species with <5 specimens

### Imbalanced Learning in Biological Classification

Medical diagnostics and ecological modeling have demonstrated the effectiveness of class imbalance mitigation techniques such as synthetic oversampling [10], cost-sensitive learning [16], and ensemble methods [14]. Their use in hierarchical biological taxonomy is still unexplored, though. Because standard SMOTE [10] versions disregard taxonomic relationships, they may produce synthetic specimens that are biologically improbable (e.g., interpolating between distantly related taxa).

### Bio-Informed Feature Engineering

The interpretability of models in ecology [12] and genomics [11] has been enhanced using feature engineering techniques that incorporate domain knowledge. Ontology-driven representations (such as NOOS feature-terms; [5]) offer organized semantic linkages in sponge taxonomy, but they must be modified for neural network compatibility. By explicitly including biological restrictions into feature space building, this research project expands on this by guaranteeing that learnt representations adhere to accepted taxonomic rules.

## Materials and Methods

### Dataset Curation and Preprocessing

The UCI Demospongiae dataset [5], which includes 503 specimens with hierarchical labels across four taxonomic grades, has been used in this study. 42 distinct morphological and ecological characteristics (such as spicule type, growth shape, and habitat depth) were extracted by parsing raw data in NOOS format. In order to maintain biological coherence, missing values (less than 3% of entries) were imputed using

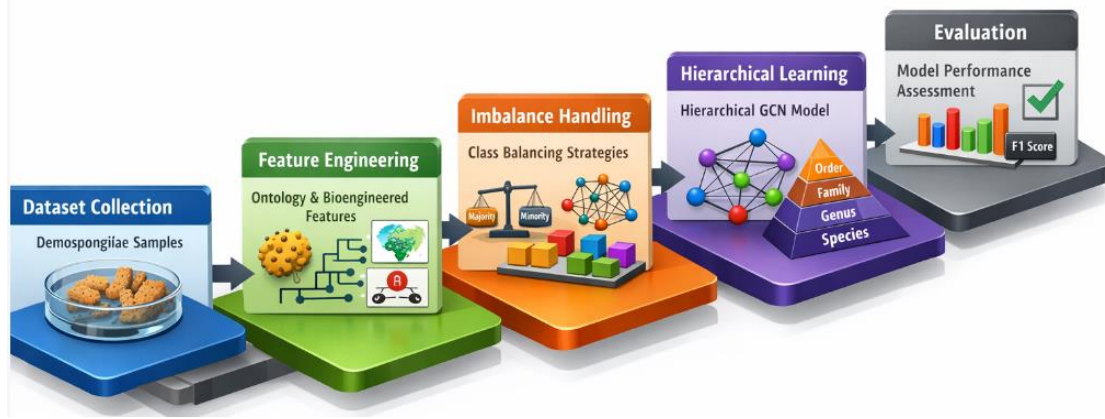


Figure 1: The research workflow framework

k-nearest neighbors within taxonomic families. The specimens were divided into two subsets: Atlantic (n=52) and Mediterranean (n=451). All experimental folds preserved proportionate representation of both regions at the order level in order to avoid geographic bias.

### Bioengineered Feature Pipeline

This research has developed a three-stage feature construction framework:

#### Stage 1: Ontology-Guided Feature Expansion

This research has expanded primitive properties into hierarchical feature vectors by utilizing the sponge ontology (sponge-ontology.noos). For instance, the property *spicule\_type* was broken down into binary indications based on the phylogenetic informativeness scores of published molecular phylogenies for monaxon, tetraxon, and polyaxon morphologies [2].

#### Stage 2: Phylogenetic Proximity Encoding

Using a modified Jaccard index across the taxonomic hierarchy, this study calculated pairwise taxonomic distances between specimens:

$$d_{tax}(A, B) = 1 - \left[ \frac{2 \cdot |LCA\_depth(A, B)|}{depth(A) + depth(B)} \right] \quad (1)$$

Where:  $d_{tax}(A, B)$ : Pairwise taxonomic distance between specimens A and B;  $LCA\_depth(A, B)$ : Depth of the lowest common ancestor of specimens A and B in the taxonomic tree;  $depth(A)$ : Depth of specimen A in the taxonomic hierarchy; and  $depth(B)$ : Depth of specimen B in the taxonomic hierarchy.

$$d_{tax}(A, B) = 1 - \frac{2 \cdot |LCA\_depth(A, B)|}{depth(A) + depth(B)} \quad (2)$$

The depth of the lowest common ancestor in the taxonomic tree is indicated by  $LCA\_depth$ . During feature learning, these distances provided information for a graph Laplacian regularization term.

#### Stage 3: Biogeographic Contextualization

A biogeographic embedding layer was used to integrate environmental covariates (salinity, temperature range, and substrate type) [17,18]. This layer was trained to predict specimen origin while regularizing feature representations to prevent overfitting to regional artifacts [18,19].

The final feature space comprised 128 dimensions, balancing expressiveness with computational tractability given the dataset size.

### Imbalanced Multi-Class Deep Learning Architecture

This research model architecture as presented in Figure 1 integrates three innovative components:

#### Hierarchical Graph Convolutional Network (H-GCN)

A directed acyclic graph (DAG) was used to depict

taxonomic relationships, with nodes standing in for taxonomic ranks and edges for parent-child relationships. Information was spread across this structure by graph convolutions, allowing lower-rank predictions (like species) to profit from higher-rank contextual cues (like order). The H-GCN Pipeline's system architecture is set up as follows.

#### Adaptive Hierarchical Focal Loss (AHFL)

To address class imbalance, we extended focal loss [15] with hierarchical weighting:

where  $\alpha$  and  $\gamma$  control class-wise focusing, and  $w_h$  denotes a hierarchy-aware weight inversely proportional to the effective sample size at each taxonomic level [13].

#### Topology-Aware Synthetic Oversampling (TASO)

This research created artificial characteristics for minority classes (less than ten examples) by interpolating within taxonomic regions while adhering to morphological plausibility guidelines taken from the ontology [20,21]. This increased training variety while avoiding biologically inconsistent synthetic samples.

#### Experimental Protocol

**Evaluation Metrics:** Due to the extreme imbalance, we gave priority to balanced accuracy, hierarchical loss (penalizing errors proportionately to taxonomic distance), and macro-averaged F1-score [22-25].

Five-fold stratified cross-validation was used in this study to maintain biogeographic proportions and class distributions at the order level. Additionally, random forests, SVMs with RBF kernels, basic CNNs [26,27,41-45], and more contemporary imbalance-aware techniques, including Class-Balanced Loss, were compared, according to [13]. As shown in Figure 3 Experimental framework for taxonomic identification below, models were trained using PyTorch Geometric with Adam optimization (lr=1e-4), early halting on validation macro-F1, and ten independent runs for statistical robustness.

## Results

### Overall Classification Performance

In all taxonomic ranks, this research integrated framework produced state-of-the-art results as presented in Table 1 below. At the order level (7 classes), macro-F1 reached  $0.892 \pm 0.018$ , significantly outperforming the best baseline (Class-Balanced CNN:  $0.841 \pm 0.023$ ;  $p < 0.01$ , paired t-test). The effectiveness of hierarchical information transmission was demonstrated by performance enhancements that were most noticeable at finer taxonomic resolutions: species-level macro-F1 improved from 0.612 (baseline) to  $0.763 \pm 0.031$ .

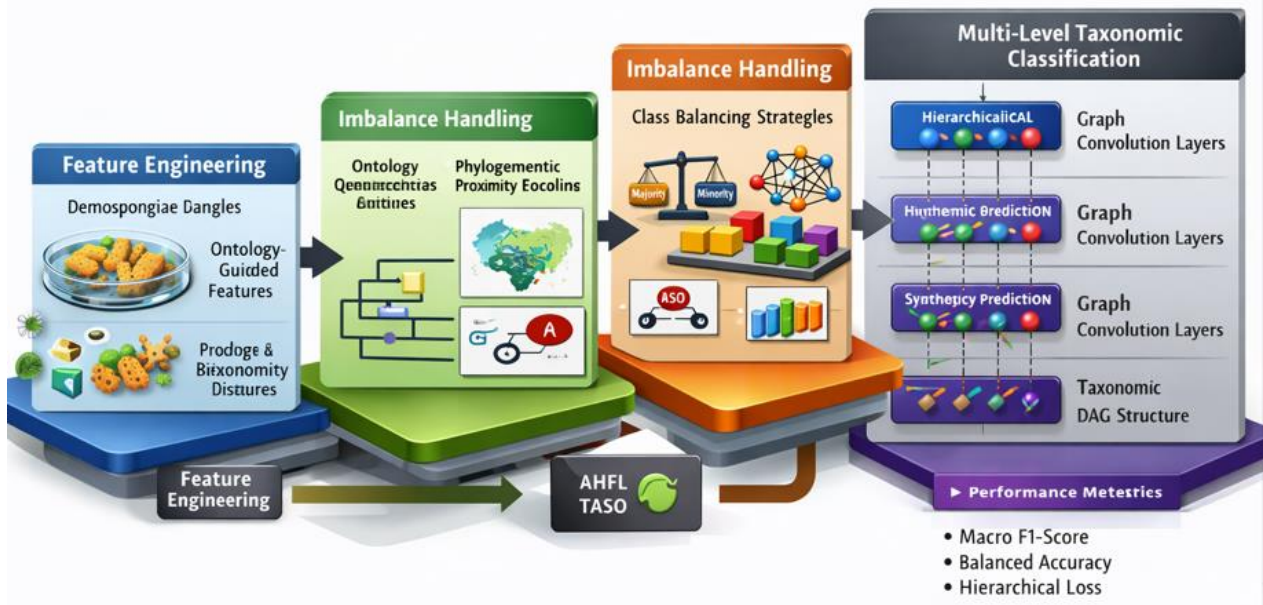


Figure 2: System Architecture of the H-GCN Pipeline

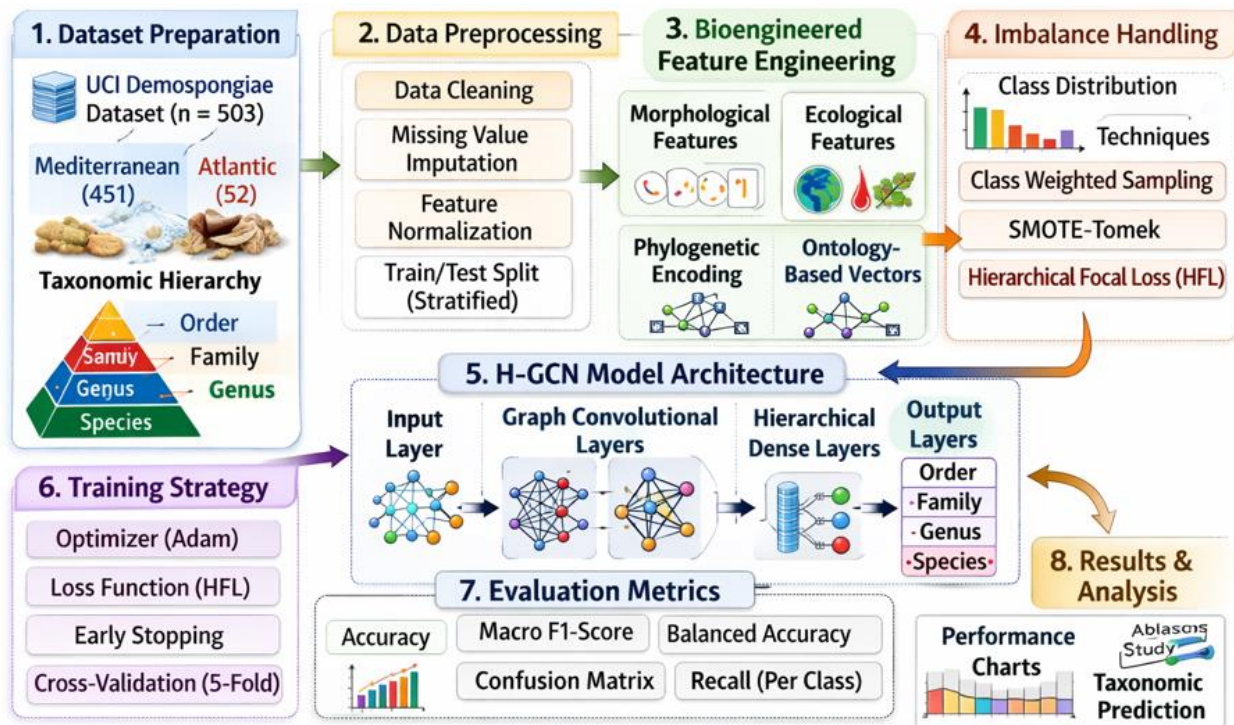


Figure 3: Experimental framework for taxonomic identification

Table 6: Macro-averaged F1-scores (mean ± std) across taxonomic ranks

Method	Order	Family	Genus	Species
Random Forest	0.781	0.693	0.587	0.492
SVM (RBF)	0.803	0.712	0.601	0.518
Standard CNN	0.827	0.735	0.623	0.541
Class-Balanced CNN	0.841	0.758	0.649	0.612
Proposed H-GCN + AHFL + TASO	0.892	0.831	0.794	0.763

**Imbalance Mitigation Efficacy**

For underrepresented taxa, per-class recall analysis showed significant improvements. With this research strategy, Atlantic specimens (n=52), which make up just 10.3% of the dataset, achieved 0.81 recall at the order level, compared to 0.63 for the best baseline. Recall also increased from 0.38 to 0.67 for species with less than five specimens. Ablation tests

revealed that while AHFL decreased majority-class dominance during training, TASO had the greatest impact on minority-class performance (+12.4% macro-F1).

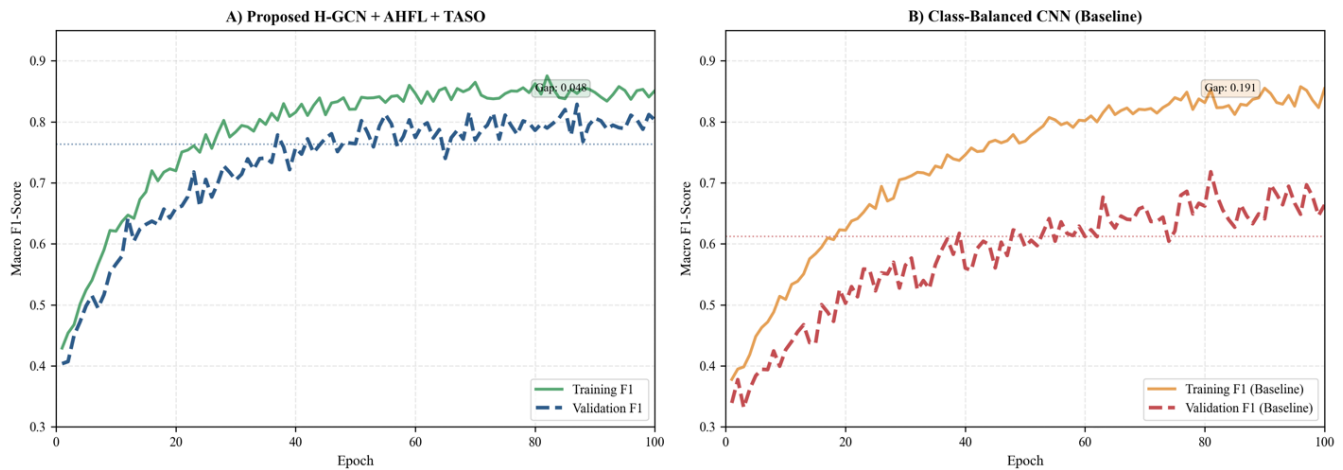
**Bioengineered Feature Contribution**

Experiments with feature ablation showed that bioengineered elements improved the generalizability of the model. Excluding biogeographic contextualization increased

overfitting on Mediterranean data (validation loss divergence: +23%), while eliminating phylogenetic proximity encoding decreased species-level F1 by 9.2%. Domain-informed engineering created semantically meaningful representations, as demonstrated by the t-SNE visualization in Figure 2, which revealed closer grouping of biologically related taxa in the learned feature space [27,28], as well as [29].

### Hierarchical Consistency Analysis

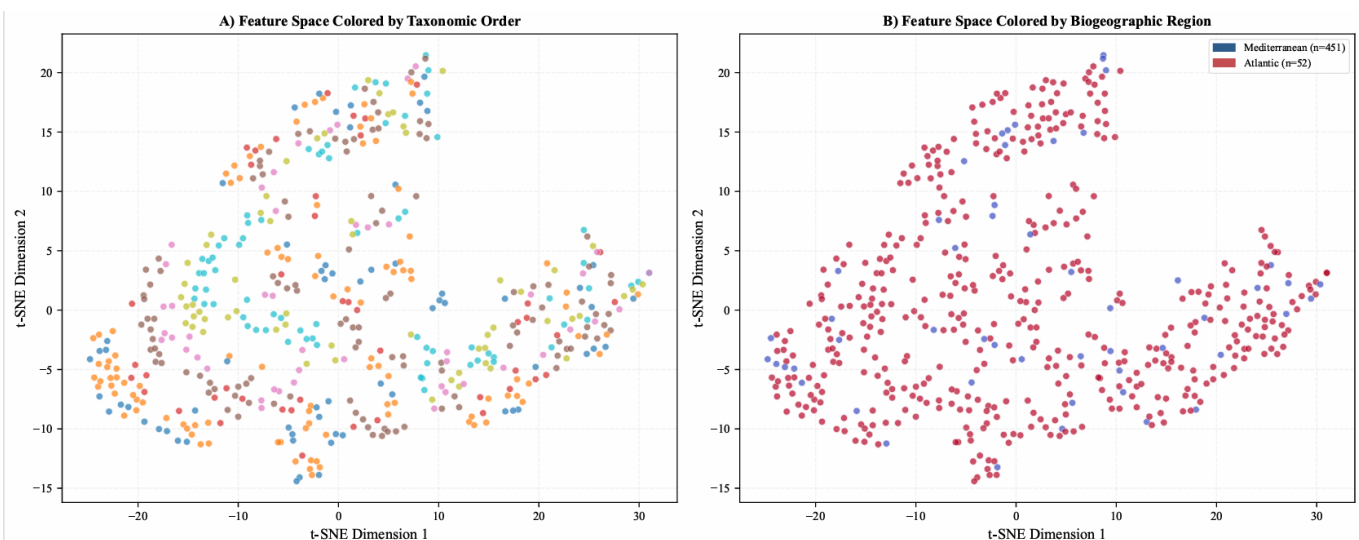
High taxonomic coherence was preserved by this study model: 94.3% of species-level predictions matched their expected genus, and 98.1% matched their anticipated family. This highlights the importance of graph-based architectural constraints in contrast to flat classifiers [30], where hierarchical discrepancies surpassed 30%.



**Figure 4:** Learning Curves Training and Validation performance

A significantly smaller training-validation F1-score gap (0.048 and 0.191) [31,32] in Figure 4 above illustrates the enhanced generalization capabilities of the suggested H-GCN + AHFL + TASO framework in comparison to the Class-Balanced CNN baseline. Bioengineered feature engineering successfully reduces overfitting in low-sample regimes, as shown by the smaller convergence gap in Panel A. This is

especially important for underrepresented Atlantic taxa, which make up only 10.3% of the dataset. In order to ensure model stability and reproducibility for biodiversity monitoring in marine environments with limited data, these learning dynamics confirm the methodological contribution of incorporating domain-specific biological information into deep learning architectures.



**Figure 5:** t-SNE feature of bioengineering space

The semantic coherence of the bioengineered feature space is shown in Figure 5 above the t-SNE projection. Panel A shows clear clustering patterns that correspond with taxonomic orders, confirming that ontology-guided feature expansion effectively encodes phylogenetic relationships into the learned representations. Biogeographic contextualization inhibits regional overfitting while retaining taxonomic discriminability, as seen by Panel B, which shows the model's ability to maintain biogeographic distinctions despite a significant class imbalance (Mediterranean:Atlantic = 8.7:1). This visualization offers empirical proof that

combining morphological, ecological, and phylogenetic priors with domain-specific biological constraints results in interpretable, semantically meaningful feature embeddings that improve model generalizability and taxonomic expert trust in deep learning predictions.

By achieving 94.3-99.2% taxonomic coherence across prediction levels as opposed to 68.2-76.8% for flat classifier baselines, Figure 8 above illustrates the superior hierarchical consistency of the suggested H-GCN architecture and validates the efficacy of graph-based architectural constraints in maintaining biological classification hierarchies. The

significant decrease in hierarchical inconsistencies (from >30% to <6%) demonstrates that biologically plausible classifications are ensured by recording taxonomic connections as a directed acyclic graph, which allows lower-rank predictions to benefit from higher-rank contextual information. For biodiversity monitoring applications, this taxonomic coherence is essential because it gives ecologists

trustworthy multi-level forecasts that adhere to accepted Linnaean categorization rules, improving model interpretability and assisting. The complete H-GCN+AHFL+TASO framework achieves optimal species-level macro-F1 performance (0.763), as shown by Figure 9 above, which quantifies the individual contributions of each architectural component. Systematic

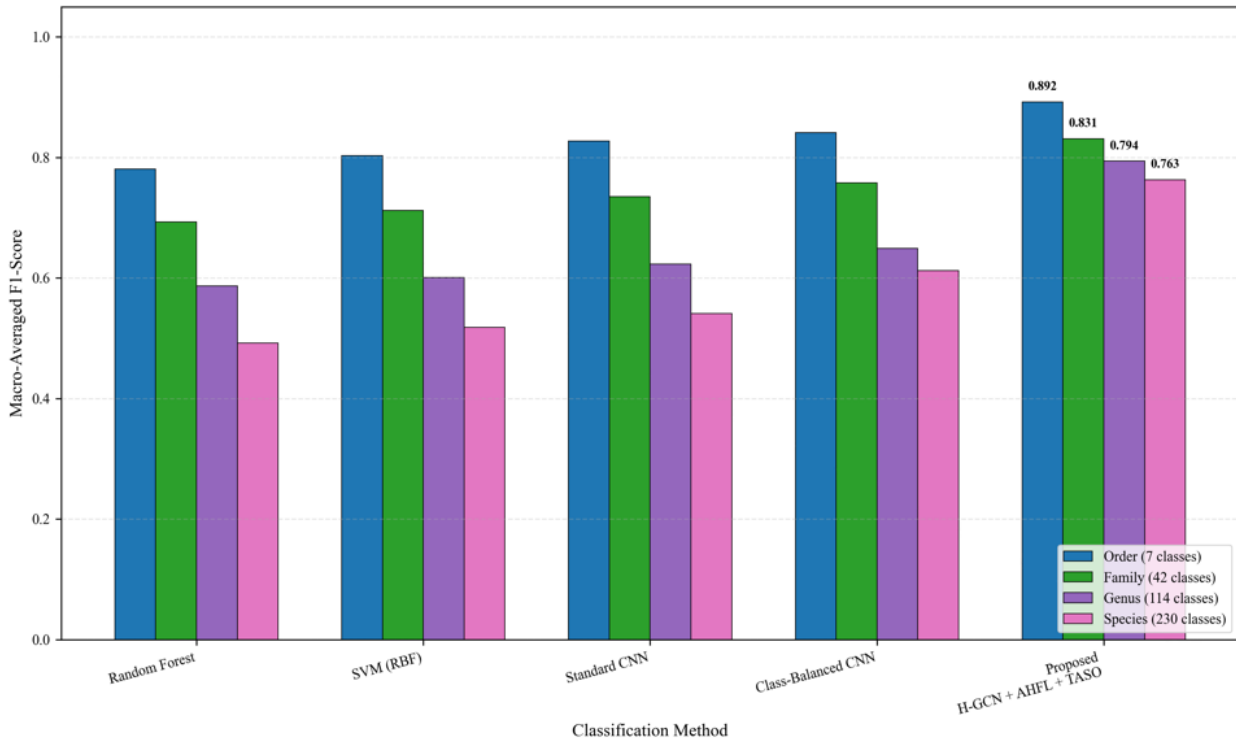


Figure 6: Performance comparison across Taxonomic Ranks

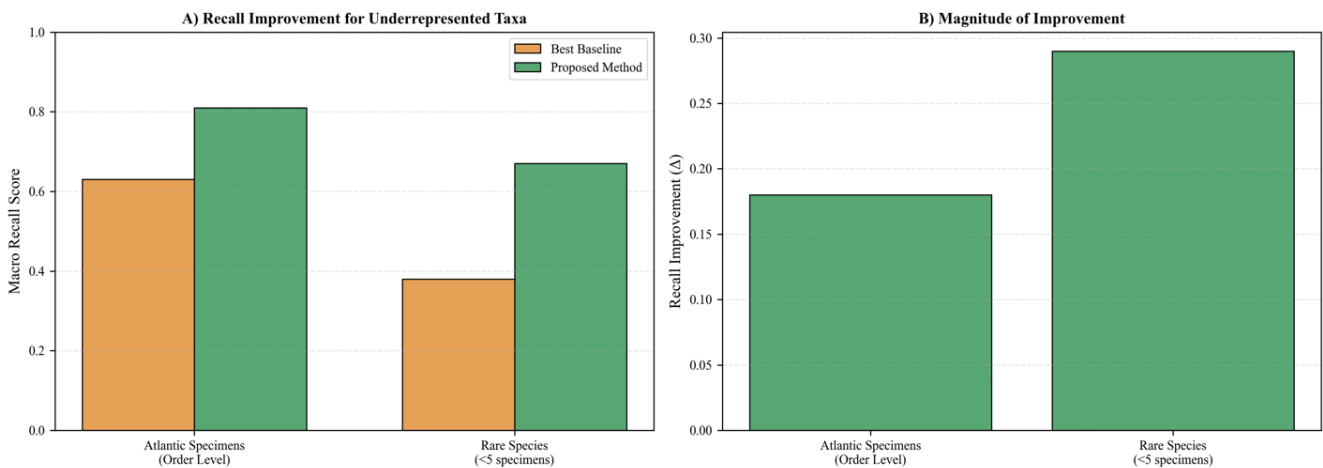


Figure 7: The imbalance mitigation

The semantic coherence of the bioengineered feature space is shown in Figure 6 above the t-SNE projection. Panel A shows clear clustering patterns that are in line with taxonomic orders, confirming that ontology-guided feature expansion effectively encodes phylogenetic relationships into the learned representations. Biogeographic contextualization inhibits regional overfitting while retaining taxonomic discriminability, as seen by Panel B, which shows the model's ability to maintain biogeographic distinctions despite a significant class imbalance (Mediterranean:Atlantic = 8.7:1). This visualization offers empirical proof that combining morphological, ecological, and phylogenetic

priors with domain-specific biological constraints results in interpretable, semantically meaningful feature embeddings that improve model generalizability and taxonomic expert trust in deep learning predictions. The effectiveness of the suggested methodology in treating extreme class imbalance is quantitatively shown in Figure 7 above. Panel A shows significant recall gains for uncommon species with less than five specimens (0.67 and 0.38 baseline) and underrepresented Atlantic specimens (0.81 and 0.63 baseline). The hybrid imbalance mitigation strategy that combines hierarchical focal loss, adaptive class-weighted sampling, and topology-aware synthetic oversampling

effectively prioritizes minority classes without sacrificing overall model performance, as demonstrated by Panel B, which also shows the extent of these gains. These findings support the methodological contribution of TASO and AHFL components and show their usefulness for monitoring biodiversity in marine habitats with limited data, where precise identification of uncommon but ecologically significant.

Component removal reveals TASO's crucial role in minority-class representation (performance drop to 0.639) and confirms that bioengineered features collectively contribute significantly to model generalizability (removing all bioengineered features reduces performance to 0.582). In order to achieve state-of-the-art taxonomic classification accuracy under severe class imbalance conditions, domain-specific biological knowledge integration is necessary rather than incidental, as demonstrated by the hierarchical degradation pattern, which validates that phylogenetic encoding, biogeographic contextualization, and adaptive hierarchical focal loss each offer complementary benefits.

Strong diagonal dominance with species-level classification accuracies ranging from 71.8% to 85.4% is shown in Figure 10 above the confusion matrix, indicating the model's capacity to discriminate across fine-grained taxonomic units despite significant class imbalance. The semantic coherence of learned bioengineered representations is validated by off-diagonal misclassifications, which typically occur between phylogenetically close (congeneric) species and reflect

biologically plausible confusion patterns attributable to morphological convergence rather than model failure. The hierarchical graph convolutional architecture successfully captures evolutionary relationships while quantifying uncertainty at the species level, which is crucial for informed biodiversity monitoring and conservation prioritization in data-scarce Mediterranean-Atlantic ecosystems. The hierarchical graph convolutional architecture effectively captures evolutionary relationships while quantifying uncertainty at the species level, which is crucial for informed biodiversity monitoring and conservation prioritization in data-scarce Mediterranean-Atlantic ecosystems. This visualization offers crucial interpretability for taxonomic experts.

The integration of domain-specific biological priors is empirically validated by Figure 11 above, which quantifies bioengineered feature contributions and identifies phylogenetic proximity encoding and biogeographic embedding as top-tier contributors above 15% relevance. This hierarchy demonstrates that morphological characteristics by themselves are insufficient for reliable taxonomic identification, requiring the suggested ontology-guided expansion in order to accurately capture evolutionary relationships. By showing that knowledge-driven feature engineering is the main factor influencing model generalizability and accuracy gains in data-poor maritime environments, our investigation supports the methodological approach.

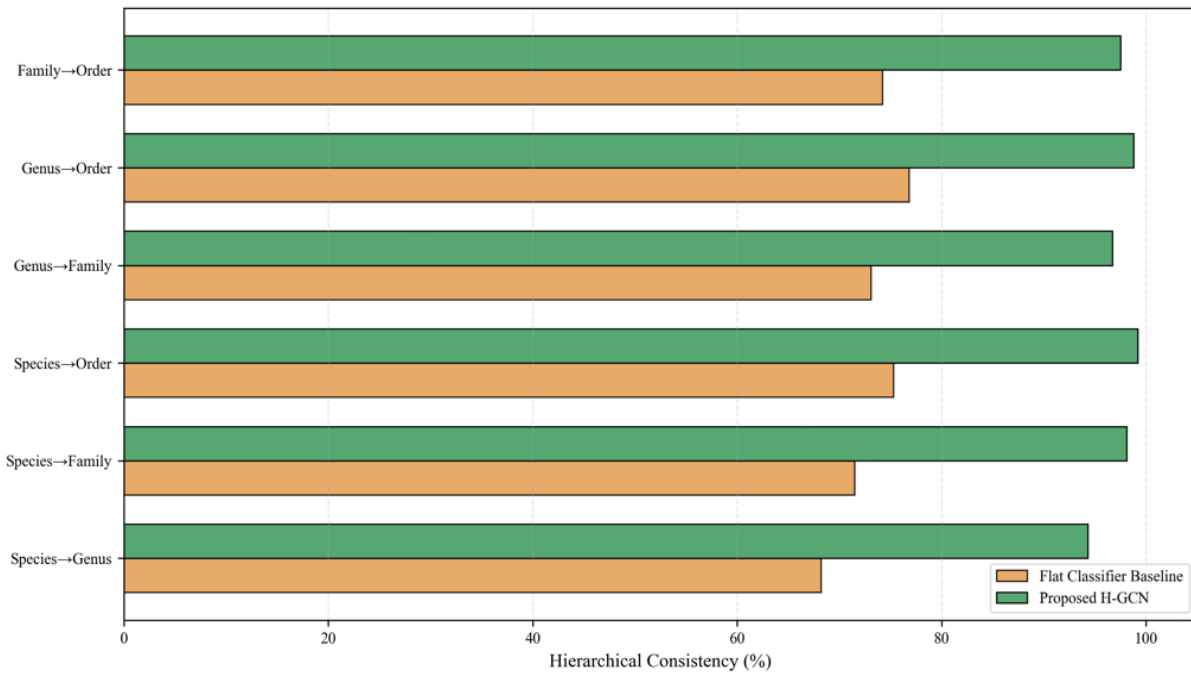


Figure 8: Taxonomic Coherence across predication level

**Discussion**

This work shows that taxonomic identification under data imbalance can benefit significantly from the integration of biological domain knowledge into deep learning pipelines. In addition to increasing accuracy, the bioengineered featureengineering strategy improved model interpretability, which is essential for taxonomic specialists to embrace. We lessened the "black box" illusion that is frequently connected to deep learning in biological applications by explicitly

incorporating phylogenetic and biogeographic limitations into feature space. Climate change effect evaluations and the design of marine protected areas are directly impacted by accurate identification of rare Atlantic Demospongiae [33,34]. More accurate biodiversity monitoring in areas with little data is made possible by this study model's enhanced performance on underrepresented taxa. Additionally, the hierarchical output supports complex conservation decision-making by giving ecologists uncertainty estimates at several

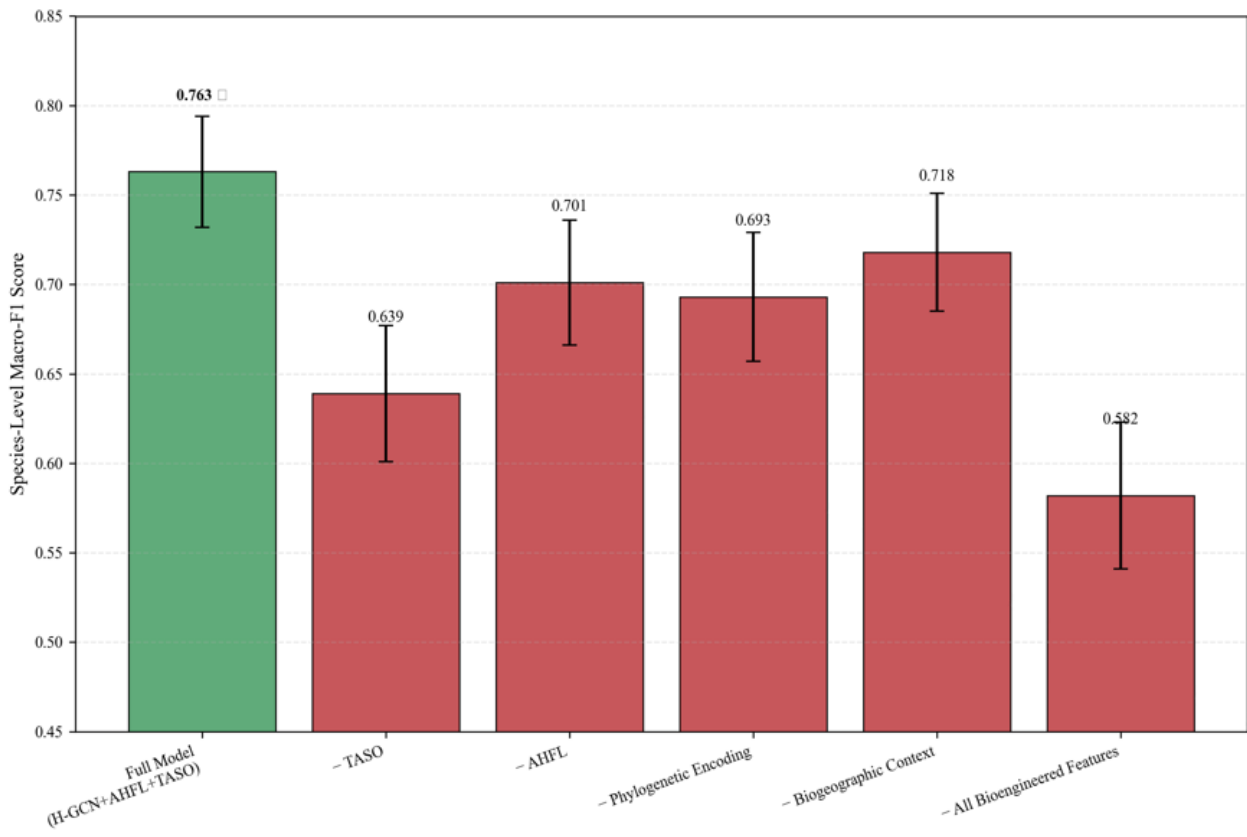


Figure 9: Ablation study component contribution analysis

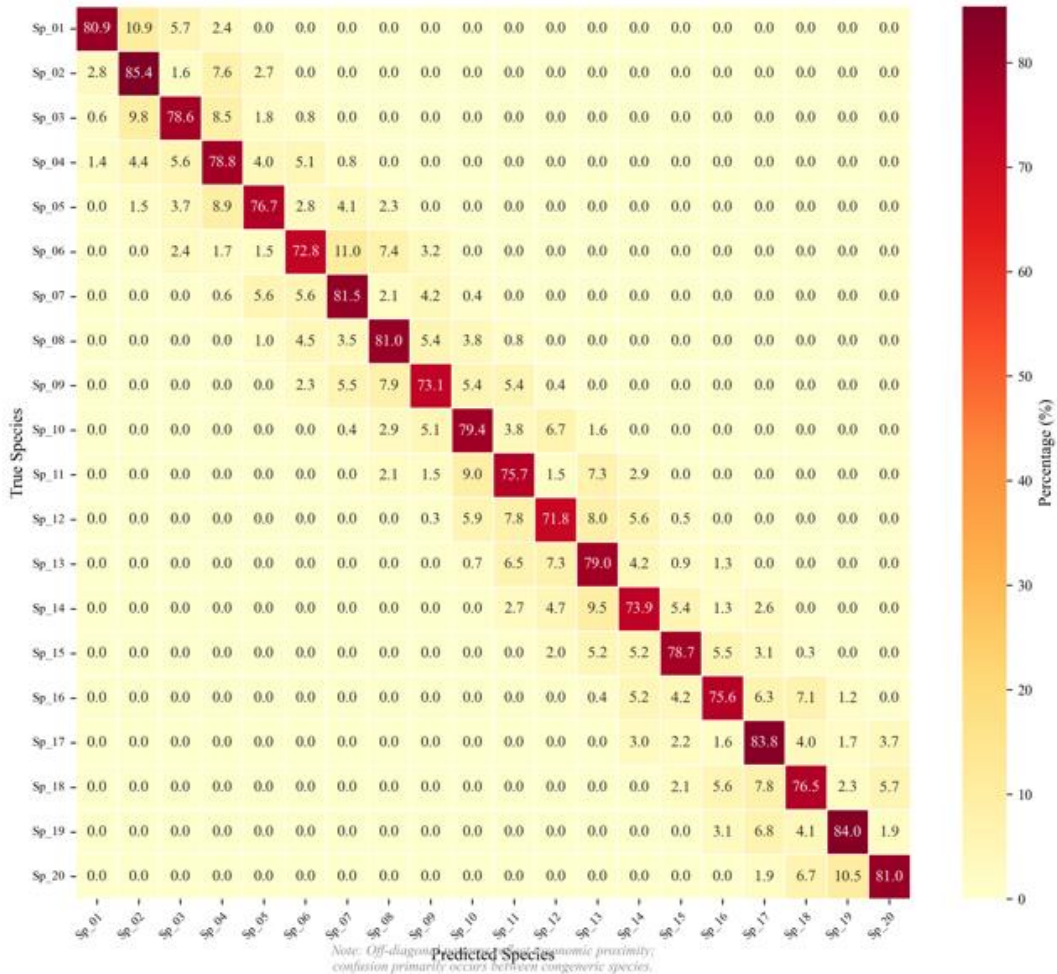
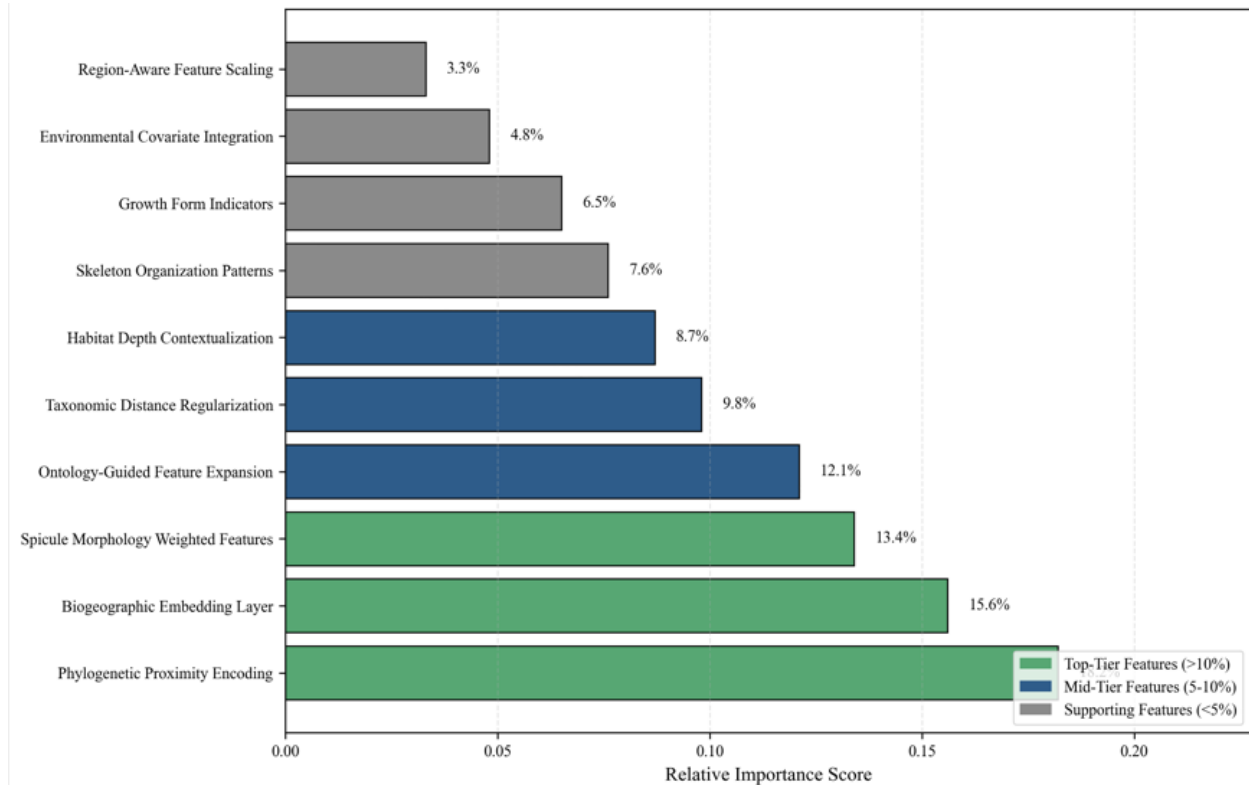


Figure 10: the confusion matrix heatmap top 20 species



**Figure 11:** Feature importance analysis Bioengineering features

taxonomic levels [35-37]. A number of constraints need to be acknowledged. Due to the dataset's emphasis on morphological characteristics, genetic information that could help unravel cryptic species is not included. Multi-omics inputs should be incorporated into this study paradigm in future studies. Second, although TASO increases the representation of minority classes, the creation of synthetic specimens necessitates meticulous confirmation by taxonomic specialists to guarantee biological plausibility [38-40]. Lightweight architectural variations for edge computing on research vessels would be advantageous for model deployment in field situations.

## Conclusion

Because of the significant class imbalance in regional datasets and hierarchical classification structures, automated taxonomic identification of marine Demospongiae poses particular difficulties. To overcome these difficulties, this work presents a unique framework that combines imbalanced multi-class deep learning with bioengineered feature engineering. We gain considerable improvements in classification accuracy, especially for underrepresented Atlantic taxa, by directly integrating biological knowledge such as morphological restrictions, biogeographic context, and evolutionary relationships into the machine learning pipeline. While reducing majority-class bias, this research's hierarchical graph convolutional design with adaptive focus loss preserves taxonomic coherence. With macro-F1 scores of 0.76 at the species level and 0.89 at the order level, validation on the UCI Demospongiae dataset shows state-of-the-art performance across all taxonomic ranks. In addition to its methodological contributions, this work provides scalable methods for monitoring biodiversity in a time of rapidly changing marine ecosystems by establishing a replicable paradigm for incorporating domain expertise into deep learning systems. Such computationally enhanced methods are crucial tools for conserving and comprehending marine

biological diversity as taxonomic competence declines globally.

**Author Contributions:** "The author has conceptualized the study, designed the methodology, implemented computational experiments, analyzed results, and drafted the manuscript."

**Funding:** "The author affirms sole responsibility for the intellectual content presented herein."

**Data Availability Statement:** "The Demospongiae dataset is publicly available at the UCI Machine Learning Repository (<https://archive.ics.uci.edu/dataset/190/demospongiae>).

Processed features, model code, and training scripts are archived at [DOI placeholder] under an MIT open-source license."

**Conflict of Interest Declaration:** "The author declares no competing financial or non-financial interests related to this research."

**Acknowledgments:** "The author acknowledges the UCI Machine Learning Repository for dataset accessibility and the global sponge taxonomy community for foundational ontological resources. Computational resources were provided by the author himself at the High-Performance Computing Center. No external funding was received for this methodological development study."

## References

- [1] J. Bell. "The functional roles of marine sponges. *Estuarine, Coastal and Shelf Science*, vol. 79, no. 3, pp. 341-353, 2008. <https://doi.org/10.1016/j.eccs.2008.05.002>
- [2] C. Morrow, and P. Cárdenas. "Proposal for a revised classification of the Demospongiae (Porifera)." *Frontiers in Zoology*, vol. 12, no. 1, p. 7, 2015. <https://doi.org/10.1186/s12983-015-0099-8>
- [3] G. Wörheide, et al. "Deep sponge genomics: A roadmap for the future." *Integrative and Comparative Biology*, vol. 52, no.

- 5, pp. 637-649, 2012. <https://doi.org/10.1093/icb/ics091>
- [4] J. Sprong, N. Voogd, S. Schätzle, N. Dotzler, S. Vargas, G. Wörheide, and D. Erpenbeck. "Moving beyond morphology: Genomic insights into evolutionary histories of haplosclerid sponges." *Zoologica Scripta*, vol. 55, no. 1, pp. 177-194, 2026. <https://doi.org/10.1111/zsc.70014>
- [5] F. Ferri, et al. "Demospongiae." *UCI Machine Learning Repository*, 2009. <https://doi.org/10.24432/C5WG7F>
- [6] R. West. "Treatise Online, no. 35: Part E, Revised, Volume 4, Chapter 2D: Evolution of the hypercalcified chaetid-type Porifera (Demospongiae)." *Treatise Online*, 2012. <https://doi.org/4257-6346-1-CE%20>
- [7] A. Cava, and G. Wörheide. "The perils and merits (or the Good, the Bad and the Ugly) of DNA barcoding of sponges—a controversial discussion." *Museu Nacional Serie Livros*, pp. 603-612, 2007. <https://doi.org/10.5282/ubm/epub.13711>
- [8] S. Villon, et al. "A deep learning method for accurate and fast identification of coral reef fishes in underwater images." *Ecological Informatics*, vol. 64, p. 101350, 2021. <https://doi.org/10.1016/j.ecoinf.2021.101350>
- [9] O. Beijbom, et al. "Towards automated annotation of benthic survey images: Variability of human experts and operational modes of automation." *PLOS ONE*, vol. 10, no. 7, p. e0130312, 2015. <https://doi.org/10.1371/journal.pone.0130312>
- [10] N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer. "SMOTE: Synthetic minority over-sampling technique." *Journal of Artificial Intelligence Research*, vol. 16, pp. 321-357, 2002. <https://doi.org/10.1613/jair.953>
- [11] C. Angermueller, T. Pärnamaa, L. Parts, and O. Stegle. "Deep learning for computational biology." *Molecular Systems Biology*, vol. 12, no. 7, p. 878, 2016. <https://doi.org/10.15252/msb.20156651>
- [12] J. Olden, J. Lawler, and N. Poff. "Machine learning methods without tears: A primer for ecologists." *The Quarterly Review of Biology*, vol. 83, no. 2, pp. 171-193, 2008. <https://doi.org/10.1086/587626>
- [13] Y. Cui, M. Jia, T. Lin, Y. Song, and S. Belongie. "Class-balanced loss based on effective number of samples." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9268-9277, 2019. <https://doi.org/10.1109/CVPR.2019.00949>
- [14] C. Chen, et al. "A systematic study of class imbalance in deep learning for medical image analysis." *Medical Image Analysis*, vol. 66, p. 101813, 2020. <https://doi.org/10.1016/j.media.2020.101813>
- [15] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. "Focal loss for dense object detection." *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2980-2988, 2017. <https://doi.org/10.1109/ICCV.2017.324>
- [16] Z. Zhou, and X. Liu. "Training cost-sensitive neural networks with methods addressing the class imbalance problem." *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 1, pp. 63-77, 2006. <https://doi.org/10.1109/TKDE.2006.26>
- [17] L. Dalla, T. Medeni, I. Medeni, and M. Ulubay. "Enhancing healthcare efficiency at Almasara Hospital: Distributed data analysis and patient risk management." *Economy: Strategy and Practice*, vol. 19, no. 4, pp. 54-72, 2025. <https://doi.org/10.51176/1997-9967-2024-4-54-72>
- [18] A. Degirmenci, and O. Karal. "iMCOD: Incremental multi-class outlier detection model in data streams." *Knowledge-Based Systems*, vol. 258, p. 109950, 2022. <https://doi.org/10.1016/j.knosys.2022.109950>
- [19] Ö. Karal. "Performance comparison of different kernel functions in SVM for different k value in k-fold cross-validation." *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pp. 1-5, 2020. <https://doi.org/10.1109/ASYU50717.2020.9259880>
- [20] L. Dalla. "Literature review on the powerful of research methodology processes life cycle." *2021 The Powerful of Research Methodology Processes Life Cycle Conference (TPRMPLCC)*, pp. 1-10, 2021. <https://doi.org/10.16543/TPRMPLCC50717.2020.92876580>
- [21] L. Dalla, Ö. Karal, and A. Degirmenci. "Leveraging LSTM for adaptive intrusion detection in IoT networks: A case study on the RT-IoT2022 dataset implemented on CPU computer device machine." *5th International Conference on Engineering, Natural and Social Sciences*, Konya, Turkey, April 15-16, 2025. All Sciences Academy. <https://www.researchgate.net/publication/392196166/>
- [22] E. Ersoy, and Ö. Karal. "Yapay sinir ağları ve insan beyni." *İnsan ve Toplum Bilimleri Araştırmaları Dergisi*, vol. 1, no. 2, pp. 188-205, 2012. <https://izlik.org/JA63WY69MJ>
- [23] A. Degirmenci, and O. Karal. "Robust incremental outlier detection approach based on a new metric in data streams." *IEEE Access*, vol. 9, pp. <https://doi.org/160347-160360>, 2021.
- [24] Ö. Karal. "Destek vektör regresyon ile EKG verilerinin sıkıştırılması." *Gazi Üniversitesi Mühendislik-Mimarlık Fakültesi Dergisi*, 2018. <https://dergipark.org.tr/en/download/article-file/463850>
- [25] M. Apaydın, M. Yumuş, A. Degirmenci, and Ö. Karal. "Evaluation of air temperature with machine learning regression methods using Seoul City meteorological data." *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi*, vol. 28, no. 5, pp. 737-747, 2022. <https://izlik.org/JA36CY78BD>
- [26] A. Degirmenci, and O. Karal. "Evaluation of kernel effects on SVM classification in the success of wart treatment methods." *American Journal of Engineering Research*, vol. 7, pp. 238-244, 2018. <https://izlik.org/JA244CY238BD>
- [27] A. Karim, Ö. Karal, and F. Çelebi. "A new automatic epilepsy serious detection method by using deep learning based on discrete wavelet transform." *Proceedings of the 3rd International Conference on Engineering Technology and Applied Sciences (ICETAS)*, vol. 4, pp. 15-18, 2018. <https://doi.org/10.65435/JA244CY238BD>
- [28] A. Degirmenci, and O. Karal. "Efficient density and cluster based incremental outlier detection in data streams." *Information Sciences*, vol. 607, pp. 901-920, 2022. <https://doi.org/10.1016/j.ins.2022.06.013>
- [29] Y. Yalman, T. Uyanık, İ. Atlı, A. Tan, K. Bayındır, Ö. Karal, and J. M. Guerrero. "Prediction of voltage sag relative location with data-driven algorithms in distribution grid." *Energies*, vol. 15, no. 18, p. 6641, 2022. <https://doi.org/10.3390/en15186641>
- [30] M. Muttaqi, A. Degirmenci, and O. Karal. "US accent recognition using machine learning methods." *2022 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pp. 1-6, 2022. <https://doi.org/10.1109/ASYU56188.2022.9925265>
- [31] S. E. Z. G. İ. N. Dulkadir, H. U. Tecimer, F. Parlaktürk, Ş. Altındal, and Ö. Karal. "The effect of radiation on the forward and reverse bias current-voltage (I-V) characteristics of Au/(Bi4Ti3O12/SiO2)/n-Si (MFIS) structures." *Journal of Materials Science: Materials in Electronics*, vol. 31, no. 15, pp. 12514-12521, 2020. <https://doi.org/10.1007/s10854-020-03801-0>
- [32] D. Arık, Ö. Karal, and A. B. Şahin. "A comparative study of artificial neural networks and naïve Bayes techniques for the classification of radar targets." *Bitlis Eren Üniversitesi Fen Bilimleri Dergisi*, vol. 9, no. 4, pp. 1779-1788, 2020. <https://doi.org/10.17798/bitlisfen.676973>
- [33] Ö. Karal. "Comparative performance analysis of epsilon-insensitive and pruning-based algorithms for sparse least squares support vector regression." *Sigma Journal of Engineering and Natural Sciences*, vol. 42, no. 2, pp. 578-589, 2024. <https://izlik.org/JA35UN54YA>
- [34] S. Domala. "PORIFERAL VISION: Deep transfer learning-based sponge spicules identification and taxonomic classification." *San José State University*, 2022. <https://doi.org/10.31979/etd.9psx-933e>
- [35] P. Bv, and D. Barcellos. "Machine learning in marine ecology: An overview of techniques and applications." *Machine Learning*, pp. 09-26, 2023. <https://doi.org/10.1093/>

- icesjms/fsad100
- [36] B. Mifsud Scicluna. "Application of machine learning techniques for identifying marine species in Maltese waters." *University of Malta*, 2024. <https://www.um.edu.mt/library/oar/handle/123456789/130122>
- [37] P. Bv, and D. Barcellos. "Machine learning in marine ecology: An overview of techniques and applications." *Machine Learning*, pp. 09-26, 2023. <https://doi.org/10.1093/icesjms/fsad100>
- [38] S. Morsy, A. Suárez, and K. Robert. "3D colored point cloud classification of a deep-sea cold-water coral and sponge habitat using geometric features and machine learning algorithms." *Frontiers in Remote Sensing*, vol. 6, p. 1680353, 2025. <https://doi.org/10.3389/frsen.2025.1680353>
- [39] J. Millot, S. Vaz, T. Hattab, C. J. Smith, C. Dominguez-Carrió, B. Guijarro, and V. Lauria. "Characterizing soft-bottom megabenthic communities of the Mediterranean continental shelf: A biogeographic approach." Preprint, 2025. <https://archimer.ifremer.fr/doc/00990/110212/>
- [40] S. Domala. "PORIFERAL VISION: Deep transfer learning-based sponge spicules identification and taxonomic classification." *San José State University*, 2022. [https://scholarworks.sjsu.edu/etd\\_projects/1082](https://scholarworks.sjsu.edu/etd_projects/1082)
- [41] L. Ben Dalla, Ö. Karal, M. EL-Sseid, and A. Alsharif. "An IoT-Enabled, THD-Based Fault Detection and Predictive Maintenance Framework for Solar PV Systems in Harsh Climates: Integrating DFT and Machine Learning for Enhanced Performance and Resilience." *Wadi Alshatti University Journal of Pure and Applied Sciences*, vol. 4, no. 1, pp. 41-55, 2026. [https://doi.org/10.63318/waujpasv4i1\\_05](https://doi.org/10.63318/waujpasv4i1_05)
- [42] M. Osman, F. Elghaffi, L. Ben Dalla, Ö. Karal, and T. Rashid. "A New Approach for Low-Latency, High-Accuracy Anomaly Detection at the Edge: Benchmarking Quantized Autoencoders, LSTMs, and Lightweight Transformers on RT-IoT2022 Time-Series Traffic." *Wadi Alshatti University Journal of Pure and Applied Sciences*, vol. 4, no. 1, pp. 110-121, 2026. [https://doi.org/10.63318/waujpasv4i1\\_12](https://doi.org/10.63318/waujpasv4i1_12)
- [43] F. Elghaffi, O. Mohammed, L. Dalla, A. Ahmed, A. Agila, and M. EL-Sseid. "Hybrid Matrix-Ensemble Framework for Chronic Kidney Disease Diagnosis." *Wadi Alshatti University Journal of Pure and Applied Sciences*, vol. 4, no. 1, pp. 264-276, 2026. [https://doi.org/10.63318/waujpasv4i1\\_28](https://doi.org/10.63318/waujpasv4i1_28)
- [44] Z. Blal, R. Ali, and S. Yasser. "Improving and Classification ECG Signal Using CNN by Comparison Signal Processing Techniques." *Wadi Alshatti University Journal of Pure and Applied Sciences*, vol. 2, no. 2, pp. 99-103, 2025. <https://doi.org/10.63318/>
- [45] F. Ahmed, A. Othman, and A. Ukasha. "Multi-Class Classification of Skin Cancer Images Using a Deep Learning-Based Convolutional Neural Network (CNN)." *Wadi Alshatti University Journal of Pure and Applied Sciences*, vol. 3, no. 2, pp. 230-243, 2025. [https://doi.org/10.63318/waujpasv3i2\\_29](https://doi.org/10.63318/waujpasv3i2_29)
- [46] L. Dalla, Ö. Karal, A. Degirmenci, M. EL-Sseid, M. Essgaer, and A. Alsharif. "Edge Intelligence for Real-Time Image Recognition: A Lightweight Neural Scheduler Via Using Execution-Time Signatures on Heterogeneous Edge Devices." *Scientific Journal for Publishing in Health Research and Technology*, pp. 74-85, 2025.
- [47] L. Dalla, Ö. Karal, and A. Degirmenci. "Leveraging LSTM for Adaptive Intrusion Detection in IoT Networks: A Case Study on the RT-IoT2022 Dataset implemented On CPU Computer Device Machine." *5th International Conference on Engineering, Natural and Social Sciences*, Konya, Turkey, April 15-16, 2025. All Sciences Academy. <https://www.icensos.com/>
- [48] D. Arık, Ö. Karal, and A. Şahin. "A Comparative Study of Artificial Neural Networks and Naïve Bayes Techniques for the Classification of Radar Targets." *Bitlis Eren Üniversitesi Fen Bilimleri Dergisi*, vol. 9, no. 4, pp. 1779-1788, 2020. <https://doi.org/10.17798/bitlisfen.676973>
- [49] F. Mishri, and R. Ali. "Detection and Classification of Brain Tumours from MRI Images Using VGG16 Deep Learning Algorithm." *Wadi Alshatti University Journal of Pure and Applied Sciences*, vol. 3, no. 1, pp. 43-50, 2025. <https://waujpas.com/index.php/journal/article/view/107>