http://www.egejfas.org
Ege Journal of Fisheries and Aquatic Sciences, 43(4), 383-392 (2026)

Su Ürünleri Dergisi (2026)
DOI: 10.12714/egejfas.43.4.07

RESEARCH ARTICLE ARAŞTIRMA MAKALESİ

Estimation of fish feeding groups, dietary breadth and trophic levels based on open data through the integration of TroPhish and FishBase data

TroPhish ve FishBase verilerinin birleştirilmesiyle balıkların beslenme grupları, besin genişliği ve trofik düzeylerinin açık veri temelli tahmini

Ertan Karahanlı

Department of Biology, Graduate School of Natural and Applied Sciences, Giresun University, Giresun, Türkiye

E-mail adress: karahanliertan46@gmail.com

Received date: .02.06.2026 *Accepted date: 22.09.2026*

How to cite this paper:
Karahanlı, E. (2026).Estimation of fish feeding groups, dietary breadth and trophic levels based on open data through the integration of TroPhish and FishBase data. *Ege Journal of Fisheries and Aquatic Sciences, 43*(4), 383-392. https://doi.org/10.12714/egejfas.43.4.07

Abstract: TroPhish investigated how well freshwater fish feeding data could be integrated with Fish-base's species, ecology, diet, and nutrient component tables. It also explored how open datasets downloaded from Fish-base could be used to model fish feeding groups, feeding diversity, and trophic levels. For the analyses, 54,750 raw feeding records and 4,571 sample-level specified "fish ID" units were used as the basis. Fish-base matching was successfully completed for 88.6% of samples and 84.2% of species. This research led to the addition of "SpecCode" and "FamCode" to 4,051 samples, feeding-based trophic levels to 3,280 samples, and nutrient data to 3,440 samples for analysis. In the "Species Group K-Fold" validation, the Random Forest model was applied and enhanced with Fish-base datasets. In the feeding group classification, the trophic identification scenario achieved 0.515% accuracy, 0.397% balanced accuracy, and a 0.390% macro F1 score. In the training data, the model's generalizability to uncharacterized families was poor. The low performance in the "Family Code Group K-Fold" validation of the same scenario is evidenced by this. While the relational power of the feeding breadth regression was found to be limited (maximum R^2 = 0.092), the trophic level regression based on fish-based data yielded realistic findings in the taxon-trait scenario (R^2 = 0.403). The results suggest that the misleading signal related to fish feeding habitat is amplified by the inclusion of TroPhish and fish- based data. Model outputs are not an automated feeding classification system that produces definitive results. They should be considered as descriptive findings generated from research-driven modeling and publicly available data.

Keywords: Fish feeding ecology, trophic level, TroPhish, fish- base, machine learning, open data integration

Öz: TroPhish tatlı su balığı beslenme verilerinin Fish-base'in tür, ekoloji, diyet ve besin bileşeni tablolarıyla ne kadar iyi entegre edilebileceğini incelemiştir. Ayrıca, Fish- base'den indirilen açık veri kümelerinin balık beslenme gruplarını, beslenme çeşitliliğini ve trofik seviyeleri modellemek için nasıl kullanılabileceğini de incelenmiştir. Analizler için temel olarak 54.750 ham beslenme kaydı ve 4.571 örnek düzeyinde belirtilen "balık kimliği" birimi kullanıldı. Örneklerin %88,6'sı ve türlerin %84,2'si için fish-base eşleştirmesi başarıyla tamamlandı. Bu araştırma, analiz için 4.051 örneğe "SpecCode" ve "FamCode", 3.280 örneğe beslenmeye dayalı trofik seviyeler ve 3.440 örneğe besin verilerinin eklenmesine yol açtı. "Tür Grubu K-Katlama" doğrulamasında, Rastgele Orman modeli uygulandı ve Fish-base veri kümeleriyle geliştirildi. Beslenme grubu sınıflandırmasında trofik tanımlama senaryosu %0,515 doğruluk, %0,397 dengeli doğruluk ve %0,390 makro F1 puanı elde etti. Eğitim verilerinde, modelin karakterize edilmemiş ailelere genellenebilirliği zayıf çıktı. Aynı senaryonun "Aile Kodu Grubu K-Katlama" doğrulamasındaki düşük performans bunun kanıtı olarak gösterilmiştir. Beslenme genişliği regresyonunun ilişkisel gücünün sınırlı olduğu (maksimum R^2 = 0,092) bulunurken, balık bazlı verilere dayalı trofik seviye regresyonu takson-özellik senaryosunda gerçekçi bulgular verdi (R^2 = 0,403). Sonuçlar, balık beslenme habitatı ile ilgili yanıltıcı sinyalin TroPhish ve balık temelli verilerin dahil edilmesiyle güçlendirildiğini göstermektedir. Model çıktıları, kesin sonuçlar üreten otomatik bir besin sınıflandırma sistemi değildir. Bunlar, araştırma odaklı modelleme ve kamuya açık verilerden üretilen tanımlayıcı bulgular olarak değerlendirilmelidir.

Anahtar kelimeler: Balık beslenme ekolojisi, trofik düzey, TroPhish, fish- based, makine öğrenmesi, açık veri entegrasyonu

INTRODUCTION

Understanding the diet of fish in freshwater ecosystems is one of the most important ways to comprehend the functioning of the ecosystem. This is because the feeding habits and related relationships of fish provide important information about energy flow, nutrient cycling, interspecies connections, and community formation (Power, 1990). The food consumed by a fish determines not only its food preferences but also its habitat, available prey, physical characteristics, life stage, and position in the food chain (Whitfield et al., 2022). Feeding data is used in a wide range of applications, from studies on environmental conditions and habitat assessments to predator-prey relationships, investigation of the effects of aggressive species, and fisheries management (Triay-Portella et al., 2022). Data on fish feeding are often obtained from different studies and using different sampling methods. As a result, data can differ in terms of both size (research sales) and taxonomic detail (Morey et al., 2024). For example, evaluating stomach contents in terms of volume, weight, abundance, or frequency makes direct comparison studies difficult. These differences also make it difficult to find a common and reliable target variable factor for modeling. TroPhish is an important data source because it brings together nutritional archives for freshwater fish reported in numerous studies. Yet, to better understand fish feeding patterns, these annals must be careful in combination with the biological and ecological characteristics of the species.

Fish- base presents a complete reserve that complements the TroPhish data in this respect. The database covers detailed data on the taxonomy, habitat, body features and trophic levels

© Published by Ege University Faculty of Fisheries, İzmir, Türkiye

of species (Froese and Pauly, 2026). In this study, fish- base was not used as a supernumerary for field data sets; rather, it was used to supplement the TroPhish records with evidence on species, family, environment, body size, alive environment and trophic level (Ridgway and Wenergway, 2025).

Machine learning is utilised in ecological studies due to its facility to understand multi-mutable and non-linear relationships (Pedregosa et al., 2011; Pichler and Hartig, 2023). However, achieving a high level of model performance on environmental datasets does not unavoidably suggest strong environmental generalisation. As model results are influenced by factors such as phylogenetic relationships amid species, taxonomic nearness, geographic sample bias, study design, as well as the lack of habitat data and alike factors, they must be inspected with great care. Also, the main theme and objective of this study is not to prove that the trophic group classification model is ultimate and generally applicable; on contrary, it is to determine in which info groups within open trophic datasets carry informative signals (Lozano-Fernandez, 2022).

The main purpose of this study is to integrate the TroPhish and Fish- base open datasets and to inspect the extent to which fish feeding collection classifications, Shannon-based dietary breadth and fish- based trophic level values can be predicted using these datasets in conjunction with variables such as habitat, morphology, geography and taxonomy. In terms of analysis of the degree of generalizability beyond potential taxonomic information and family level structure, model validation methods grouped at the species level and at the family code scale were also likened.

MATERIALS AND METHODS

Merging data sources

In this study, two open-access data sources were used: TroPhish and Fish- base. TroPhish contains feeding records for different freshwater fish species compiled from various studies. The dataset includes information such as fish species, the taxon or class of prey consumed, the percentage of the diet, the sampling characteristics, habitat, fish size and geographical location. Parquet tables containing the main characteristics, ecology, diet and food items consumed by the species were obtained from Fish-base.

The TroPhish records were first conserved in their raw form and then summarised so that each “fish- id” represented a distinct specimen, making them suitable for modelling. The matching between TroPhish and Fish- base was carried out using species names and “SpecCode” information. Afterward, Fish- base’s tables relating to ecology and diet were linked to the relevant specimens via “SpecCode” and/ or “DietCode”. In this way, the dietary information obtained from TroPhish for each “fish-id” was combined with the taxonomy, morphology, habitat and trophic level variables found in Fish- base inside a single data assembly (Table 1).

Table 1. Scope of the dataset and results of the comparison with FishBase

Analytical level	Number	Description
Raw diet record	54.750	TroPhish raw diet-record structure
Sample/ fish id level	4.571	Main observation unit used in modelling
Samples matched with Fish- Base SpecCode	4.051	Samples successfully matched at species/ fish- Base level
Species matched with Fish- Base	826	Matched subset of TroPhish species
Samples with Fish- Base diet trophic level	3.280	Samples with trophic level summary from the diet table
Samples with food items information	3.440	Samples linked to the food items table

Creating target variables

Three target variables were used for modelling. The first target was the feeding guild class derived according to the dominant prey group. Prey classes were converted into more interpretable ecological groups: diets dominated by insects were coded as insectivore; diets dominated by fish or vertebrates were coded as piscivore/ vertebrate feeder; and diets dominated by crustaceans or zooplankton were coded as crustacean/ zooplankton feeder. Other groups were arranged under herbivore/ algivore, detritivore, benthic invertebrate and mixed/ other categories.

The second target variable was a Shannon- based diet breadth indicator calculated from diet percentages for each fish id. Although this indicator provides a useful summary of diet diversity, it is sensitive to the resolution of reported prey categories, sampling intensity and study design. For this reason, diet breadth regression was evaluated as a supportive modelling output rather than as the main prediction target.

The third target variable was trophic level summarized from the fish- base diet table. Unlike the TroPhish- derived feeding guild model, this target was used to test the predictability of a fish- based numerical indicator of trophic position.

Involvement variable star and model situations

The model involvements have been grouped under three chief information layers: habitat and morphology variables sourced from TroPhish; taxonomic, morphological and habitat variable star obtained from Fish- base; and trophic and analytic variables also based on fish- base. From the TroPhish dataset, the variable quantity used were fish dimension, habitat type, latitude, longitude, absolute latitude and sampling month. From Fish- base, variables connecting to genus, FamCode, length, DemersPelag, attendance in freshwater, brackish water and saline environments, depth variety and ecological environments were included in the model.

The study likened three different model scenarios. Whilst Scenario A comprised only habitat and morphologic variables resultant from TroPhish, Scenario B also combined taxonomic, morphological and habitat characteristics sourced from Fish-base. Scenario C, in addition to these variables, included diagnostic variables derived from Fish- base's trophic level and nutrient tables. Mainly with respect to the classification of feeding groups, Scenario C was evaluated not as a self-regulating field prediction model, nevertheless as a diagnostic sympathy scenario representative the influence of Fish- base's trophic information layers to classification achievement (Table 2).

Table 2. Model scenarios and interpretation framework

Scenario	Name	Main variables	Interpretation
A	TroPhish habitat- morphology	Length, habitat, location and month	Independent baseline open-data scenario
B	Fish- Base taxon- trait	A + Genus, FamCode and Fish- base morphology/ habitat fields	Scenario enriched with taxonomy and species traits
C	Fish- Base trophic- diagnostic	B + FishBase diet/ ecology/food items trophic summaries	Scenario testing the diagnostic contribution of trophic information layers

Model development and performance evaluation

The performance of Random Forest and Logistic Regression models in classifying feeding groups was compared. Feeding diversity and trophic level values were obtained from Fish-base. Ridge and Elastic Network regression were used for prediction with this data. Missing data were addressed using appropriate preprocessing methods. Categorical variables were labeled to make them suitable for modeling. The machine learning process was performed using the scikit-learn library (Pedregosa et al., 2011). The Random Forest algorithm was implemented in accordance with the general approach outlined by Schonlau (2023), that is, it was used as an ensemble learning method combining the outputs of multiple decision trees.

Instead of a rigid, random row-based approach, a combined, enhanced cross-screening method was used. To reduce species-specific memorization in the model, a species group k-folding technique was added. This technique ensures that comparison data between similar species are not presented together across the test and application range. Fish-based Family Code grouping approximates the Family Code group k-folding time. This allows for a more conservative fit and a more comprehensive refinement of the model, reducing taxonomic flows outside of family boundaries.

RESULTS

FishBase pairing success and the distribution of feeding groups

The analysed TroPhish dataset contains 54.750 raw feeding records and 4.571 fish- id units identified at the specimen level. At the sample level, 4,051 of these were successfully matched with Fish-base data, resulting in an 88.6% match percentage. At the species level, 826 species were matched with Fish-base, yielding an 84.2% success rate. The complete and accurate match rates demonstrate that the main modeling procedure was able to consistently integrate the additional information provided by Fish-base. According to the results of the distribution of feeding groups at the species level, insectivores and crustaceans/zooplankton feeders were found more frequently in the dataset than benthic invertebrates and those categorized as "unknown/other". Because of this imbalance between classes, model performance was evaluated not only in terms of overall accuracy, but also considering macro F1 and balanced accuracy metrics (Table 3).

Table 3. Distribution of dietary groups identified at species level

Feeding guild	Number of species	Share (%)
Insectivore	288	29.358
Crustacean/ zooplankton	203	20.693
Herbivore/ algivore	151	15.392
Piscivore/ vertebrate	148	15.087
Detritivore	82	8.359
Unknown/ other	66	6.728
Benthic invertebrate	26	2.650
Other animal	17	1.733

Model performance regarding the classification of dietary groups

The inclusion of Fish- base data in the model led to more successful results in the classification of trophic groups compared with the baseline scenario, which only included habitat and morphology variables sourced from TroPhish. Under species group k- fold validation, the best result was obtained from the C- Fish- base- trophic- diagnostic scenario using the Random Forest model; accuracy, balanced accuracy and macro F1 were calculated as 0.515, 0.397 and 0.390 (Figure 1).

In addition, the results of validating the same scenario using the family code group k- fold process, which showed a decline in balanced accuracy and macro F1 scores, indicated that the model's capacity to generalise beyond taxonomic patterns at the family level remains limited. This finding suggests that taxonomic and trophic variables contribute to the prediction process; however, it also suggests that some of the information provided may stem from the similarity between species within the same family.

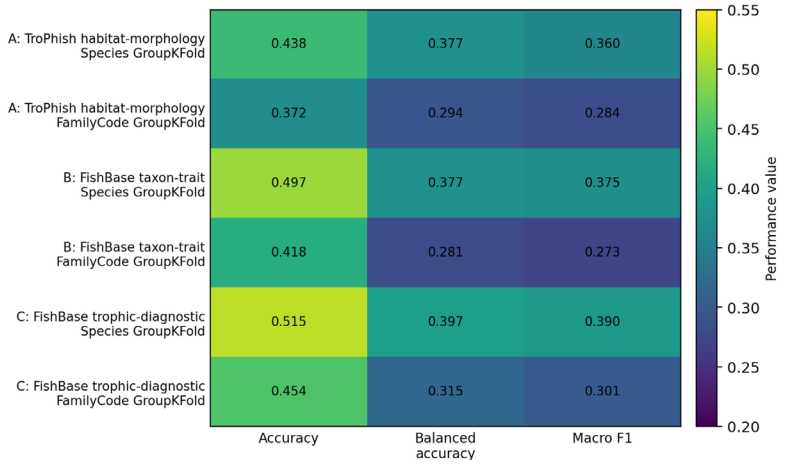

Figure 1. Model in the classification of dietary groups

Class-level performance and confusion matrix

The model was able to differentiate between insectivorous and fish-eating/vertebrate-eating groups more reliably, according to class-level analyzes and evaluations. On the other hand, for fewer distinguishable groups in the sample, including those that consume benthic invertebrates, prediction accuracy was shown to be quite poor. The main trophic groups' comparable environments and numerical differences between classes are thought to be the cause of this (Table 4). An inspection of the misperception medium exposed important classification errors, particularly between crustacean/ zooplankton feeders, insectivores and fish/ vertebrate feeders. It is therefore more appropriate to interpret the model productions not as a means of transmission ultimate dietary group labels to species, but as indicators tightfitting which groups can be illustrious more dependably in open data sources (Figure 2).

Table 4. Performance results of the most successful dietary group model by class

Class	Precision	Recall	F1- score	Support
Benthic invertebrate	0.013	0.007	0.009	141
Crustacean/ Zooplankton	0.512	0.410	0.455	1.032
Detritivore	0.305	0.321	0.313	215
Herbivore/ Algivore	0.469	0.504	0.486	462
Insectivore	0.566	0.637	0.600	1.446
Mixed/ Other	0.460	0.163	0.241	386
Piscivore/ Vertebrate	0.545	0.738	0.627	832

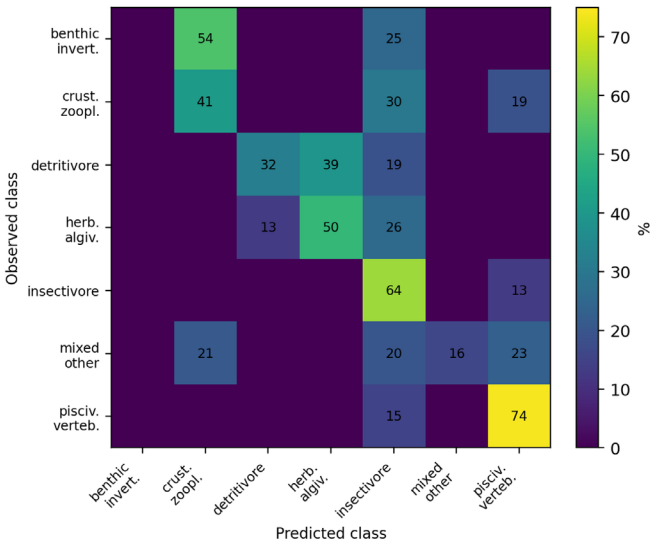

Figure 2. Row-wise normalised confusion matrix (%) for the most successful dietary group model

Regression results relating to dietary diversity

The highest R^2 value in the estimation of dietary breadth was obtained by applying the Elastic Net model under the Species group k- fold validation for Scenario C, which was enriched with fish- based data (R^2 = 0.092). This consequence, which signifies only an incomplete development ended the baseline scenario, designates that the model can clarify only a small proportion of the difference in dietary breadth, with the mainstream of the variation remaining mysterious (Table 5).

The distribution of observed and predicted values revealed that model predictions were concentrated around the mean values and did not adequately represent the lower and upper extremes of the dietary breadth. This finding suggests that the dietary breadth indicator can be significantly influenced by factors such as local species diversity, sampling resolution, reporting detail and study design (Figure 3).

Table 5. Performance of regression models for approximating dietary variety

S	V	M	N	R^2	RMSE	MAE
C	Sp	EN	4.514	0.092	0.422	0.353
B	Sp	EN	4.514	0.085	0.424	0.354
A	Sp	EN	4.514	0.053	0.431	0.364
A	Fam	EN	4.514	0.051	0.432	0.364
A	Sp	Ridge	4.514	0.039	0.434	0.366
A	Fam	Ridge	4.514	0.015	0.440	0.370

(Abbreviations: S, scenario; V, validation scheme; M, model; N, number of samples; Sp, species-group K-fold cross-validation; Fam, family-code-group K-fold cross-validation; EN, Elastic Net regression; R^2, coefficient of determination; RMSE, root mean squared error; MAE, mean absolute error. A, TroPhish habitat–morphology scenario; B, FishBase taxon–trait scenario; C, FishBase trophic–diagnostic scenario).

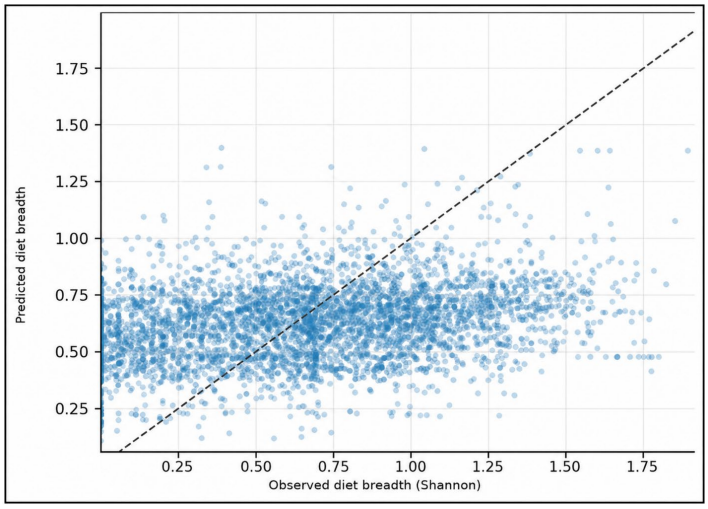

Figure 3. Distribution of experiential and forecast values in the dietary variety regression

FishBase trophic level regression

Compared to the dietary breadth model, the regression model for determining trophic level developed from Fish-base showed a better explanatory power. Using the elastic net approach under "Species Group K-Fold" validation, the "B-Fish-base-taxon-trait" scenario produced the best outcome. The R2, RMSE, and MAE values for this model were determined to be 0.403, 0.476, and 0.383, respectively (Table 6). In the "Family Code Group K-Fold" validation, the greatest R^2 value decreased to 0.231, indicating that the trophic level forecast benefited somewhat from the taxonomic and morphological similarities of species within the same family. However, the findings demonstrate that Fish- base species traits provide a more distinct, consistent and modelable signal in the approximation of numerical trophic level compared to dietary group organisation (Figures 4 and 5).

Variable importance and model interpretability

The most profitable dietary group model's variable significance values have demonstrated that the predictions are influenced by a variety of ecological factors, including morphology, taxonomy, sampling conditions, geographic features, and trophic indicators derived from Fish-base. Fish-length-cm, Genus, FamCode, n-diet-records, longitude, abs-latitude, latitude, fb-diet-Troph-mean, fb-diet-Troph-median, month, eco-FoodTroph, and Length were recognized as the variables having the highest significance in the model (Table 7 and Figure 6).

Table 6. Performance of trophic level regression models derived from Fish- base

S	V	M	N	R^2	RMSE	MAE
B	Sp	EN	3.278	0.403	0.476	0.383
B	Sp	Ridge	3.278	0.277	0.524	0.407
B	Fam	EN	3.278	0.231	0.540	0.439
A	Sp	EN	3.278	0.042	0.603	0.476
A	Sp	Ridge	3.278	0.033	0.606	0.481
B	Fam	Ridge	3.278	-0.019	0.622	0.483

(Abbreviations: S, scenario; V, validation scheme; M, model; N, number of samples; Sp, species-group K-fold cross-validation; Fam, family-code-group K-fold cross-validation; EN, Elastic Net regression; R^2, coefficient of determination; RMSE, root mean squared error; MAE, mean absolute error. A, TroPhish habitat–morphology scenario; B, FishBase taxon–trait scenario).

This finding implies that physical traits, taxonomic closeness, geographic location, and sample density all work together

to distinguish the eating patterns of fish in the dataset. However, it is not appropriate to take the variable significance levels as direct markers of causal effects. While the importance of the "n-diet-records" variable shows that sample density also supports evidence to the approximate estimate technique, the prominence of the "Genus" and "FamCode" variables indicates that the model makes use of taxonomic similarities.

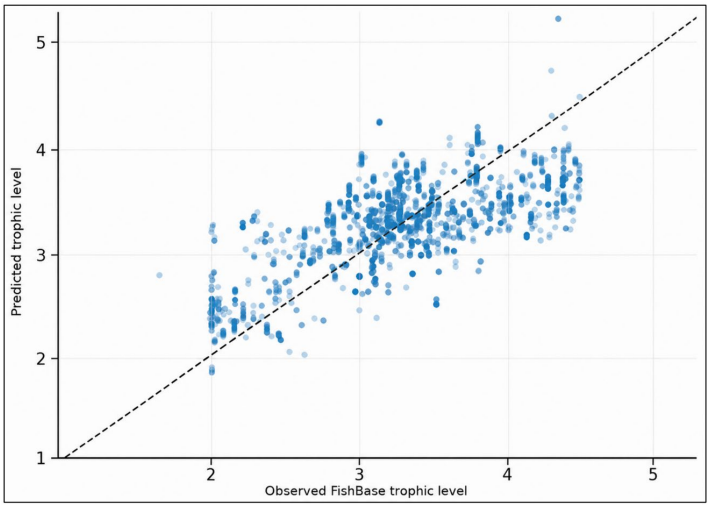

Figure 4. Distribution of predicted and assessed values in the trophic level regression based on fish- based data

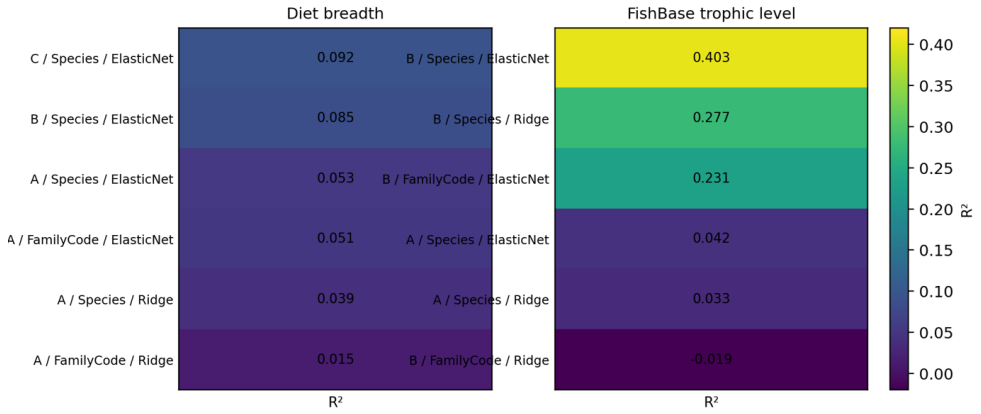

Figure 5. Comparison of R^2 values obtained from nutrient diversity and trophic level regressions with the heat map

Table 7. Significance levels of the 12 variables that stood out in the most successful nutritional group model

Group	Variable	Importance
Morphology trait	Fish length (cm)	0.073
Taxonomy	Genus	0.065
Taxonomy	FamCode	0.055
Sampling context	Diet record count	0.052
Geography	longitude	0.051
Geography	Absolute latitude	0.048
Geography	latitude	0.047
External trophic context	Fish- Base diet trophic level (mean)	0.047
External trophic context	Fish- Base diet trophic level (median)	0.045
Sampling context	month	0.042
External trophic context	Ecology food trophic level	0.041
Morphology trait	Length	0.032

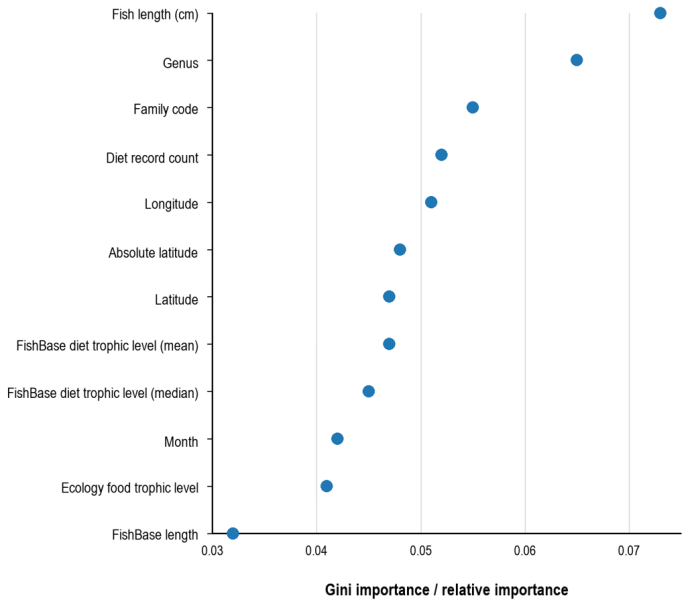

Figure 6. Dot plot representation of variable importance levels in the most successful nutrition group model

DISCUSSION

This study demonstrates that combining the open-access TroPhish and Fish-base datasets allows for preliminary simulation of fish feeding groups, feeding diversity, and trophic levels. An interpretable machine learning framework was also used for the model. The study findings show that the inclusion of Fish-base data improves the models' capacity to recognize ecological patterns, particularly in feeding group classification and trophic level estimation. The generated models should not be considered as an automated feeding classification method providing definitive results. Given the heterogeneous trophic data collected from multiple sources, the model outputs should be understood as explanatory results highlighting ecological, taxonomic, and sampling-based trends.

In previous studies, various components of fish trophic ecology have been modelled using stomach content data, morphological characteristics, phylogenetic relationships or environmental variables. However, most of these studies have been limited to a specific ecosystem type, geographical region or a single trophic response variable. As far as we are aware, this study provides the first comprehensive open-data framework designed to incorporates global freshwater fish diet data from TroPhish with Fish-base's taxonomic, morphological, ecological and trophic data sources, therefore considering feeding groups, dietary breadth estimated using the Shannon index and numerical trophic level within a unified modelling framework. The study also distinguishes itself from previous research in that it evaluates the contribution of different data layers to the model in a stepwise manner. To this end, a three-stage modelling approach was adopted, comprising a baseline scenario containing only TroPhish habitat and morphology variables, a scenario enriched with taxonomic and species-specific characteristics from Fish-base, and a scenario incorporating trophic- diagnostic variables.

Another methodological contribution of the study is its comparison of the "Species Group K- Fold" and "Family Code Group K- Fold" methods, the latter of which offers a more rigorous validation approach. Thanks to this design, it has been possible to assess both whether the integration of open data sources improves model performance and to what extent the observed performance improvement is influenced by taxonomic proximity. Thus, the study goes beyond merely identifying the model that achieves the highest prediction success; it also examines ecological signal, taxonomic dependence and the model's applicability to new taxonomic groups within the same framework.

The dietary group model that demonstrated the highest performance under species group k- fold validation achieved accuracy of 0.515, balanced accuracy of 0.397 and macro F1 of 0.390. This result demonstrates that adding taxonomic and trophic variables provided by Fish- base to the baseline model- which contained only TroPhish data- improved classification performance. Similarly, Parravicini et al. (2020) demonstrated that the trophic groupings and dietary relationships of coral reef fish can be predicted based on phylogenetic distance and maximum body length. This result is in agreement with with the high importance associated with fish length, Genus and FamCode variables in the present study. But, the markedly higher predictive performance achieved by Parravicini et al. (2020) should not be directly compared with the results of this study. That study is based on a coral reef community with more distinct ecological limitations, consistently compiled gut content data, direct phylogenetic information and a different prediction objective. The current study, on the other hand, brings together heterogeneous freshwater fish records from

different geographical regions, sampling periods and analytical approaches, and tests model performance through cross-validation based on taxonomic groups. In this respect, the study assesses the transferability of the models to different freshwater species and families under more challenging and realistic conditions.

The use of direct observational data to construct trophic groups is similar to the approach taken by Thompson et al. (2020), who utilised a comprehensive marine trophic matrix to develop a reproducible trophic group indicator that could be used to monitor environmental change. The researchers reported that trophic groups are associated with predators' body size, phylogenetic relatedness, and habitat characteristics; they also noted that data-driven classification could reduce the need for subjective assessments based on expert opinion. The present study takes that approach a step further by investigating whether groups derived from the trophic matrix can be estimated using different open-data layers. It also testing whether the classification signal distinguishing dietary groups is preserved when specific species or families are excluded from the model training.

These results can also be linked to the comprehensive study conducted by Thompson et al. (2025), which brought together 944.129 stomach samples from 227 predator species and 2.158 prey taxa from the North Atlantic and Arctic Oceans. The researchers demonstrated that predator identity, predator and prey size, and the functional composition of prey communities play significant roles in predicting species- and body-size- specific feeding features. The results of the present study also confirm the importance of these biological and taxonomic characteristics in explaining the feeding patterns of fish. However, this study differed from that of Thompson et al. (2025) in that it utilised globally compiled freshwater fish records, modelled three different trophic targets within the same framework, and tested directly whether predictive success was maintained beyond taxonomic similarities at the species and family levels.

Environmental conditions may be another source of uncertainty encountered in the classification of trophic groups. Sun et al. (2024) found that environmental variables measured in situ- mainly water depth, transparency and turbidity- accounted for 61.2 per cent of the observed variation in the biomass of fish trophic groups in the Muling River Basin. The relatively poor performance of the baseline model using TroPhish morphology and habitat inputs does not directly contradict this result. While TroPhish relies on habitat descriptors compiled from various geographic locations and studies that are heterogeneous in terms of scope and documentation structure, Sun et al. (2024) used comprehensive ecological data collected within a single river system. Habitat variables found in global open-access datasets may strongly clarify the feeding structure within a given ecosystem, but may not adequately reflect small-scale environmental conditions that define prey availability and actual feeding patterns.

The low explanatory power regarding dietary breadth is an ecologically expected result. Keppeler et al. (2025), in their study covering 541 freshwater fish populations belonging to 358 species and 82 families, showed that trophic niche breadth is associated with environmental conditions, fish diversity, trophic position, fin morphology and body size variation within the population. Moreover, the researchers' estimates were influenced by the number of individuals studied, environmental conditions and the years in which sampling took place.

The reduced performance of the dietary breadth model is also consistent with the consequences obtained by Olden et al. (2025). The researchers found that, on a global scale, there was only a weak relationship among morphological traits and the isotopic niches of freshwater fish. In contrast, sampling location accounted for a much larger proportion of the observed variation than morphological traits. This finding suggests that readily seeming species traits cannot be regarded as a complete proxy for directly measured trophic data.

Higher descriptive power is demonstrated by the regression model used to predict trophic level based on fish-based data. The fish-based taxon trait scenario produced the greatest results using the Elastic Network approach under the "Species Group K-Fold" validation; R2 RMSE and MAE values were found to be 0.403, 0.476, and 0.383, respectively. This finding suggests that, in contrast to sample-specific feeding diversity, numerical trophic level exhibits a more reliable and consistent relationship with species-level taxonomic and morphological traits. According to Keppeler et al. (2020), fish body size and trophic level have different relationships depending on their feeding strategy and the functional traits that go along with it. When various trophic strategies were taken into account collectively, no common link that applied to all fish could be found, even though a positive association is more common among carnivorous fish. Additionally, it was discovered that tooth morphology, body depth, and mouth opening all affect this association. These results validate the importance of fish length as a crucial component in the existing model. At the same time, they explain why fish with different taxonomic and functional characteristics cannot be accurately predicted by length alone.

Consequently, whilst the arithmetical trophic level seems more predictable than dietary breadth, the relations with morphological characteristics are not universal but are context-dependent and subject to specific circumstances.

In conclusion, while arithmetic trophic level appears more predictable than feeding diversity, the relationships with morphological features are not universal. These features are context-dependent and subject to specific conditions. The decline in prediction performance for feeding groups and trophic levels in the "Family Code Group K-Fold" validation suggests that some of the signal used by the model stems from taxonomic similarities between closely related groups.

According to Yates et al. (2023), the cross-validation method requires the selection of variables and methods appropriate to the specific ecological prediction problem under consideration, noting that inappropriate data processing methods can lead to the model presentation appearing higher than it actually is. While "Species-Group K-Fold" keeps similar species specimens from appearing together in both the training and test sets, "Family Code Group K-Fold" provides a more stimulating valuation, requiring the model to predict taxa fitting to relations that were never visible during the model training stage. Because of this, the model's limitations shouldn't be the only reason for the experienced decrease in performance. The results show that the feeding ecology of fish has a hierarchical taxonomic pattern and that it is cooler to foresee formation between groups of related species than to conclude to groupings that are taxonomically distinct.

The study has several limitations because the data was gathered from various sources. The extent and completeness of the information offered in Fish-base differs by species, even as TroPhish records are gathered from various geographical regions, time periods, sampling strategies, and food analysis techniques. Furthermore, the identification of dietary groups and the computation of dietary breadth based on the Shannon index may be impacted by the defining of prey categories at various levels of precision. Statistical pre-processing techniques cannot adequately account for these discrepancies, which result from ecological factors and observation methods.

In general, modeling studies that seek to clarify the feeding ecology of freshwater fish benefit greatly from the merging of TroPhish and fish-based data. This study's wider significance comes from proving that the trophic targets under investigation are not all equally predictable. The representation of classes in the dataset and the taxonomic structure have an impact on the classification of feeding groups, even if the numerical trophic level shows a more consistent association with taxonomic data and species features. Temporarily, localized ecological factors and study sampling features have been the primary determinants of nutrient diversity at the sample level. This comparison aids in identifying which trophic traits may be evocatively projected from existing open-access resources and which traits require more thorough test group data and standardized field observations.

The study's primary findings show that high prediction accuracy is attained for all goals and that adding fish-based data to the model improves its ability to detect environmental signals. Additionally, it demonstrates unequivocally that the taxonomic level at which generalization is assessed affects prediction ability. Classified information scenarios and validation methods aggregated at the species and family levels may be used to construct models, and a mix of open data can lead to performance gains. It is feasible to differentiate between performance gains brought about by the advantages of taxonomic closeness by using this approach. The approach offers a more thorough and environmentally responsible evaluation than only reporting model correctness.

In future studies, the boundaries of feeding groups should be defined more clearly by using more detailed prey taxonomy data, and methods aimed at reducing class imbalance should be applied. Furthermore, it is important to standardise variables such as life stage, size class, prey availability and sampling effort, and incorporate them into the models. Regional environmental variables such as water temperature, primary productivity, hydrological conditions and habitat complexity may contribute to the improvement of models, particularly those relating to dietary breadth.

The substitutability of models must be verified using completely independent datasets reflecting different geographical regions and sample programmes. In parallel as grouped authentication approaches, in which certain species or families are excluded from the training set, provide valuable internal evidence, they do not replace external validation. Once issues of taxonomic dependency, class imbalance and external validity have been addressed more comprehensively, more sophisticated model explainability approaches- including SHAP-based methods- can be utilised (Lundberg et al., 2020).

CONCLUSION

The modeling of fish feeding ecology utilizing open data sources has benefited greatly from the merging of TroPhish and Fish-based data. Fish-base acquired a high match rate at the specimen level; trophic level indicators, morphological and habitat features, family code, and genus have given the model balancing and descriptive layers of information.

When it came to classifying trophic groups, models enhanced with Fish-base data outperformed the baseline model, which just used the TroPhish variable quantity. However, the Family Code Group k-fold authentication's performance fall emphasizes the necessity for caution when simplifying the models to include families that aren't represented in the training data. As a result, the industrialized method should be viewed as a descriptive modeling framework that identifies taxonomic and environmental indicators in publicly available datasets rather than as a categorization tool that gives species definite trophic groupings.

Given the existing nature of the accessible data, it is challenging to consistently estimate dietary breadth, as evidenced by the regression model's low performance. The trophic level model that came from FishBase, on the other hand, has a greater explanatory capacity. This finding implies that taxonomic and morphological traits are more consistently correlated with numerical trophic rank than with dietary breadth.

ACKNOWLEDGEMENTS AND FUNDING

The author gratefully acknowledges the developers and maintainers of TroPhish and FishBase for making the datasets publicly available. The author received no external funding for this study.

AUTHORSHIP CONTRIBUTIONS

Ertan Karahanlı: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Visualization, Writing – original draft, Writing – review & editing, and Project administration. The author read and approved the final manuscript.

CONFLICT OF INTEREST

The author declares that there is no conflict of interest.

ETHICAL APPROVAL

This study utilised only publicly available secondary ecological data sources. The study did not require approval from an ethics committee, as it did not include human participants, personal data or any animal experiments carried out by the author.

DECLARATION OF AI USE

Artificial intelligence-supported tools have been used solely for the purpose of editing the text in terms of language and style, correcting grammatical and spelling errors, and improving clarity and readability. These tools have not been used for the generation or fabrication of data, the conduct of analyses, the interpretation of findings, the preparation of figures, tables or graphs, or to substitute for the author's scientific judgement. The author has checked all outputs generated by AI-assisted tools, reviewed and approved the final version of the article, and assumes full responsibility for the content of the work.

DATA AVAILABILITY

The TroPhish data used in this study are openly available in Zenodo at https://doi.org/10.5281/zenodo.13850491. FishBase data are openly accessible at https://www.fishbase.org (version 02/2026). The cleaned analysis tables, derived variables, figures and model outputs supporting the findings are available from the corresponding author, upon reasonable request.

REFERENCES

Froese, R., & Pauly, D. (2026). FishBase (Version 02/2026) [Database]. https://www.fishbase.org

Keppeler, F.W., Giarrizzo, T., Montaña, C.G., Adite, A., Akın, Ş., Angelini, R., Arantes, C.C., Benedito, E., Bokhutlo, T., El-Sabaawi, R., Garcia, A., Gonzalez-Bergonzoni, I., Hoeinghaus, D., Hoffman, J., Jensen, O.P., Jeppesen, E., Kopf, R.K., Layman, C.A., Lopez, E.O., … Winemiller, K. O. (2025). What determines trophic niche breadth? A global analysis of freshwater fishes using isospaces. *Reviews in Fish Biology and Fisheries*, *36*(1), 1–30. Article 13. https://doi.org/10.1007/s11160-025-10023-2

Keppeler, F.W., Montaña, C.G., & Winemiller, K.O. (2020). The relationship between trophic level and body size in fishes depends on functional traits. *Ecological Monographs*, *90*(4), e01415. https://doi.org/10.1002/ecm.1415

Lozano-Fernandez, J. (2022). A practical guide to design and assess a phylogenomic study. *Genome Biology and Evolution*, *14*(9), evac129. https://doi.org/10.1093/gbe/evac129

Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, *2*(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

Morey, K.C., Myler, E., Hanner, R., & Tetreault, G. (2024). Taxonomic blind spots: A limitation of environmental DNA metabarcoding-based detection for Canadian freshwater fishes. *Environmental DNA*, *6*(6), e70054. https://doi.org/10.1002/edn3.70054

Olden, J.D., Keppeler, F.W., Jardine, T.D., Boulêtreau, S., Harrod, C., Jackson, M.C., McIntosh, A.R., Perga, M.-E., South, J., Tedesco, P.A., Vagnon, C., & Cucherousset, J. (2025). Looks can be deceiving: Morphological traits are weakly associated with the isotopic niches of freshwater fishes. *Global Ecology and Biogeography*, *34*(9), e70124. https://doi.org/10.1111/geb.70124

Parravicini, V., Casey, J.M., Schiettekatte, N.M.D., Brandl, S.J., Pozas-Schacre, C., Carlot, J., Edgar, G.J., Graham, N.A.J., Harmelin-Vivien, M., Kulbicki, M., Strona, G., & Stuart-Smith, R.D. (2020). Delineating reef fish trophic guilds with global gut content data synthesis and phylogeny. *PLOS Biology*, *18*(12), e3000702. https://doi.org/10.1371/journal.pbio.3000702

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825-2830. https://jmlr.org/papers/v12/pedregosa11a.html

Pichler, M., & Hartig, F. (2023). Machine learning and deep learning—A review for ecologists. *Methods in Ecology and Evolution*, *14*(4), 994–1016. https://doi.org/10.1111/2041-210X.14061

Power, M. E. (1990). Effects of fish in river food webs. *Science*, *250*(4982), 811–814. https://doi.org/10.1126/science.250.4982.811

Ridgway, J., & Wesner, J. (2025). A global dataset of freshwater fish trophic interactions. *Scientific Data*, *12*, Article 160. https://doi.org/10.1038/s41597-024-04211-1

Schonlau, M. (2023). Random forests. In M. Schonlau (Ed.), Applied statistical learning: With case studies in Stata (pp. 183–204). Springer International Publishing. https://doi.org/10.1007/978-3-031-33390-3_10

Sun, X., Wang, K., Zhang, G., Ren, H., & Yu, H. (2024). Spatial–temporal patterns of fish trophic guilds in a freshwater river wetland ecosystem of northeastern China. *Ecology and Evolution*, *14*(7)e11711. https://doi.org/10.1002/ece3.11711

Thompson, M.S.A., Pontalier, H., Spence, M.A., Pinnegar, J.K., Greenstreet, S.P.R., Moriarty, M., Hélaouët, P., & Lynam, C.P. (2020). A feeding guild indicator to assess environmental change impacts on marine ecosystem structure and functioning. *Journal of Applied Ecology*, *57*(9), 1769–1781. https://doi.org/10.1111/1365-2664.13662

Thompson, M.S.A., Preciado, I., Maioli, F., Bartolino, V., Belgrano, A., Casini, M., Cresson, P., Eriksen, E., Hernandez-Milian, G., Jónsdóttir, I.G., Neuenfeldt, S., Pinnegar, J.K., Ragnarsson, S., Schückel, S., Schückel, U., Smith, B.E., Torres, M.Á., Webb, T.J., & Lynam, C.P. (2025). Fish functional groups of the North Atlantic and Arctic oceans. *Earth System Science Data*, *17*(6), 2447–2462. https://doi.org/10.5194/essd-17-2447-2025

Triay-Portella, R., Martín, J.A., Luque, L., & Pajuelo, J.G. (2022). Relevance of feeding ecology in the management of invasive species: Prey variability in a novel invasive crab. *Estuarine, Coastal and Shelf Science*, *274*, 107949. https://doi.org/10.1016/j.ecss.2022.107949

Whitfield, A.K., Able, K.W., Blaber, S.J.M., Elliott, M., Franco, A., Harrison, T.D., & Houde, E.D. (2022). Feeding ecology and trophic dynamics. In A.K. Whitfield, K.W. Able, S.J. M. Blaber, & M. Elliott (Eds.), Fish and fisheries in estuaries: A global perspective (pp. 255–331). Wiley. https://doi.org/10.1002/9781119705345.ch5

Yates, L.A., Aandahl, Z., Richards, S.A., & Brook, B.W. (2023). Cross validation for model selection: A review with examples from ecology. *Ecological Monographs*, *93*(1), e01557. https://doi.org/10.1002/ecm.1557