FROM MORPHOLOGY TO MACHINE LEARNING: RANDOM FOREST-BASED APPROACH FOR FISH SPECIES IDENTIFICATION

A. Zafar, M. U. Sarwar*, M. K. Hanif and M. R. Talib

Government College University, Faisalabad, Pakistan

Corresponding Author's Email: mumersarwar@gcuf.edu.pk

ABSTRACT

Fish are important components of freshwater ecosystems and are one of the main sources of protein for humans. One of the main problems in fishery resource investigation is the accurate identification of fish species and the accurate discrimination of fish stocks. In the context of large data, machine learning techniques as developing data processing techniques have gradually replaced traditional methods. This study demonstrated an automated study, using machine learning, to classify fish species by morphological features (e.g., body shape, size, and fins etc.). Machine learning-based fish species identification is essential for automated ecological monitoring, biodiversity assessment, and fisheries management. This study used a dataset of 540 samples across 20 distinct fish species, a Random Forest model was trained and evaluated using a 10-fold cross-validation strategy, outperforming alternative comparative models in classification accuracy. The procedure used to prepare the dataset consisted of preprocessing, including categorical to binary encoded categorical data (converting non-numeric data to numbers the machine could interpret). Several machine learning models were examined, with the Random Forest (RF) providing the best results, and the RF became a trained, tested, and saved machine learning model for user applications. Subsequently, a slight implementation interface was additionally developed to demonstrate the practical applicability of the trained classification framework that allowed users to enter fish features and receive predictions. The proposed framework illustrates how machine learning approaches may support species identification in fisheries and ecological studies, increase prediction accuracy, and have potential applications in fish farming and research as well as in aquatic ecosystem protection and conservation.

Keywords: Machine learning, Fish Morphology, Random forest, Streamlit web analytic Approach

https://doi.org/10.36899/JAPS.2026.5.0121 Published first online July, 15, 2026

INTRODUCTION

Freshwater ecosystems are extremely diverse and represent a small percentage of the Earth's water resources. Fish communities have a significant ecological and economic value in these ecosystems, in providing nutrient cycling, food-web stability, fisheries production and human nutrition. There is therefore a need for accurate fish species identification for biodiversity assessment, fisheries management, ecological monitoring and conservation planning. An increasing number of freshwater fishes, however, are facing risks, with a range of problems that are affecting their habitats, including some environmental changes linked to climate change. These problems are causing habitat degradation, pollution, overexploitation, hydrological alteration, and climate-related environmental changes, all of which are threats to freshwater fish populations and require reliable and scalable monitoring approaches (Rashid *et al*., 2024; Farooq *et al*., 2026). Freshwater fishes are a significant part of food security and aquatic biodiversity in South Asia, especially in the Indus River system and should be focused on the development of species identification systems that are specific to local ecological contexts (Sajjad *et al*., 2023).

Traditional methods of fish identification are largely based on manual taxonomic examination of the external features and measurements of the fish. While these methods are still basic in fisheries biology and ichthyology, they are often time consuming, labor intensive, and require specialist taxonomic expertise. Identification is often challenging in the field especially when the species are closely related and have similar external characteristics, or a large number of samples need to be processed under different environmental conditions. Traditional methods of monitoring, such as manual sorting and net sampling, can also lead to observer variation and bias in data collection (Jalal *et al*., 2020; Wang *et al*., 2021). This has led to the growing focus on automated and semi-automated classification methods that are able to help ecological monitoring and fisheries research in a more efficient way in recent years (Mohammadisabet *et al*., 2025).

Ghouri *et al*. (2020) reported that there are 531 different kinds of fish in Pakistan; of which 233 live in freshwater and 298 live in saltwater. Economically, 78 out of the 233 freshwater fish species are considered more significant. The River Indus is home to over 180 different types of fish (Sheikh *et al*., 2017; Khan *et al*., 2021). In total, 86 fish species have been identified as important in Pakistan, including 8 exotic and 78 native species. These species are valued for their endemism, economic importance, conservation status according to the IUCN, and rarity (Rafique and Khan, 2012). Fish are classified by scientists on the basis of their physical characteristics: shape of the body, body texture, body colour, and shape of the head (Tseng and Kuo, 2019). Monitoring the fish population in rivers and lakes is a vital step in designing a superior system to manage the ecosystem and obtain a sustainable fishery (Jalal *et al*., 2020). These species should be properly taxonomically characterized to allow different scientific, conservation, and fisheries management purposes (Mujtaba and Mahapatra, 2022). The limitations and material demands of the traditional fish identification methods which involve manual surveying and nets are daunting (Abinaya *et al*., 2022). Surveys conducted manually also are labour-intensive and need human observers to collect the data which can introduce bias and inaccuracy (Wang *et al*., 2021). Besides, an unintended consequence of net utilization is the unavoidable entrapment of non-target organisms which can make the data interpretation challenging and impede the effective monitoring of fish populations (López-Barajas *et al*., 2024). Chandak *et al.* (2023) also observed the identification process of freshwater fish using the assistance of hi-tech devices that can quickly recognize and decode species, types, habitat, and ecological functions. To address these issues, they came up with an artificial intelligence-based fish recognizer system, which relies on machine learning algorithms. Machine learning refers to mathematical models capable of executing a particular task without any command to do so. There are three broad categories of machine learning problems: unsupervised, supervised and reinforcement learning. Unsupervised learning is directed at discovering patterns in unlabeled input data, such as clusters of similar data. Reinforced learning is a technique that involves interaction with an environment and a reward/penalty system, while supervised learning finds a mapping of the input to a labeled output, such as differentiating between two categories (Kühn *et al*., 2025).

In fisheries science, the wide range of machine learning and computer vision techniques has increasingly become applicable for fisheries research, biodiversity monitoring, and aquatic ecological research. In recent years, it has been found that deep learning, convolutional neural networks and ensemble learning techniques, can be applied to automatic fish detection, classification, tracking and biomass estimation both in laboratory and field conditions (Villon *et al*., 2022; Kühn *et al*., 2025). Underwater fish recognition and ecological monitoring using computer vision-based systems have been found to be promising, especially for the integration of image databases with the real-time analysis of underwater videos (Allken *et al*., 2019; Li and Du, 2022). Likewise, recent reviews have found that machine learning frameworks are an emerging component in fisheries management, and have been shown to be beneficial for species identification, ecological monitoring and biodiversity assessments (Al-Abri *et al*., 2025). Many advanced image-based systems, however, need a huge amount of computational power, a large amount of annotated data, special imaging equipment or special environmental conditions and may not be applicable in resource-limited areas.

Morphological features are still one of the most convenient and easily accessible methods of fish species discrimination, especially in developing regions where advanced imaging systems and molecular facilities may not be readily available. The quantitative morphometric traits (body length, size of the head, fin position, body depth) can give biologically important information, which may help taxa differentiation and ecological assessment. Morphometric analysis can be done with relatively simple field equipment and standardized measurement procedures, compared to image-based systems that can be affected by underwater visibility, lighting, and image quality. It has also been recently demonstrated that machine-learning algorithms are able to make use of morphometric datasets for biological classification purposes, especially for cases in which nonlinear relationships between the anatomical variables are involved (Chandak *et al*., 2023; Monteiro *et al*., 2023). Therefore, a combination of traditional morphometric methods and machine learning may provide an economical and large scale solution for monitoring freshwater biodiversity and fisheries related applications.

The aim of the present study was to develop and test a morphology based machine learning approach using morphological data of selected river systems in Punjab, Pakistan for the classification of freshwater fish species. Freshwater ecosystems of Pakistan are very rich and harbor a great variety of economically and ecologically important fishes. The present study specifically targeted selected fish of the fresh-water type from the major freshwater rivers of Punjab like river Indus, river Chenab and river Jhelum. These river systems are ecologically significant freshwater systems and provide significant fisheries and biodiversity resources in the area. A total of 20 economically important fish species were sampled from the rivers Indus, Chenab and Jhelum and were subjected to multiple machine learning classifiers like Random Forest, Decision Tree, K-Nearest Neighbors and Support Vector Machine models. The study also investigated the feasibility of utilizing the morphological features for species identification and the performance of different classification techniques. This work does not mainly deal with the implementation of the software, it emphasizes the ecological and fisheries value of applying machine learning to support morphological classification in a regional biodiversity monitoring context. It was hypothesized that machine learning outperforms the traditional morphometric keys for fish identification in different rivers of the Punjab. The outcomes will contribute to continuing efforts aimed at integrating computational approaches into fisheries science and provide a preliminary framework for supporting morphology-based species identification.

MATERIALS AND METHODS

Data Collection: The primary dataset was obtained directly from the Department of Zoology at Government College University, Faisalabad (GCUF), as part of a PhD research project. This departmental dataset is not publicly available and was specially acquired to ensure that fish species-specific information is authentic, verified, and relevant to the ecological focus of the study. Fish samples were randomly collected from the study area by local fishermen using drag and cast nets during September 2022 to April 2024, after approval from the ethics review committee (ERC) of Government College University Faisalabad, Pakistan, vide Ref. No. GCUF/ERC/22/390. Morphometric measurements were recorded manually by trained researchers from the Department of Zoology, Government College University Faisalabad, following standard ichthyological measurement protocols. All morphometric measurements were obtained from fresh fish specimens shortly after collection to minimize preservation-related deformation. No formalin or ethanol preserved specimens were included in the morphometric analysis, so as to reduce the possibility of dimensional changes in the specimen due to the preservative. Only morphologically intact and healthy specimens were included in the dataset. Measurements were recorded according to standard ichthyological morphometric protocols using digital Vernier calipers with 0.01 mm precision and measuring scales for larger body dimensions. The sampled fish represented a moderate size range commonly observed in local fisheries, thereby reducing extreme size-related variability during morphometric analysis. To reduce observer bias and recording errors, prior to inclusion into the final data set, all morphometric measurements were carefully inspected and any inconsistencies checked for data entry. Measurement verification procedures have been put in place, but formal inter-rater reliability statistics were not computed and are suggested for application in future studies to further quantify the measurement reproducibility. Species identification was confirmed using regional taxonomic keys (Mirza, 2003) and expert verification.

Dataset Composition and Structure: The overall dataset consisted of mainly random sample collection from selected rivers and image-based records, which captured a wide range of fish species. To develop models of machine learning that are applicable and successful in any setting, it is necessary to obtain diverse and similar data on the distribution of fish species within populations, places, and the time when the samples were collected. The quality of raw data as well as the performance of machine learning models was improved with preprocessing and filtering.

Morphological Features: The current study involve 20 economically important freshwater fish species that are commonly consumed in the study area namely *Catla catla*, *Labeo calbasu*, *Labeo gonius*, *Labeo bata*, *Labeo rohita*, *Cyprinus carpio*, *Cirrhinus mrigala*, *Ompok bimaculatus*, *Cirrhinus reba*, *Salmostoma phullo*, *Systoma sarana*, *Heteropneustes fossilis*, *Eutropiichthys vacha*, *Notopterus notopterus*, *Chitala chitala*, *Clupisoma garua*, *Wallago attu*, *Sperata seenghala*, *Rita rita* and *Bagarius bagarius* (Table 1). Fish species were sampled from different locations of different rivers namely River Indus [(Mianwali (MW, 32°35′7″N 71°32′37″E), Kallur kot (KK, 32°9'23" N and 71°15'34" E), Dera Ghazi Ghat (DGG, 30°03′00″N 70°37′58.8″E), Chashma Barrage (CB, 32°24′N 71°21′E), Head Taunsa (HT, 30°54′N 70°58′E), and Head Muhammad wala (HMW, 30.2992° N, 71.3460° E)], River Chenab [(Trimmu Barrage (TB, 31.1448° N, 72.1464° E) and Chiniot Pull (CP, 31°45′10″N 72°56′53″E)] and River Jhelum [Rasul Barrage (RB, 32°40′49″N 73°31′15″E)] Table 2. From each species, morphological features including Total length (TL), Fork length (FL), Standard length (SL), Pre-dorsal length (PrDL), Pre-pelvic length (PPL), Head length (HL), Pre-pectoral length (PPecL), Preorbital length (POL), Postorbital length (PostOL), Body depth (BD), Head depth (HD), Peduncle depth (PD), Height of dorsal fin (HDF), Height of anal fin (HAF), Pectoral fin length (PFL), Pelvic-fin base length (PBL), Anal-fin base length (ABL), Eye orbit (EO), Eye diameter (ED), Postdorsal length (PsDL), Pre-anal length (PAL) and Snout length (SnL) were recorded with scale and digital calipers. In addition to the primary morphometric variables, Interorbital distance (IO), Upper jaw length (UJL), and Lower jaw length (LJL) were incorporated during the feature preparation stage because of their potential discriminatory value in species classification. The final dataset therefore included all morphometric variables consistently used during model training and validation (Figure 1, Table 3).

Figure 2 represents 20 fish species (outer ring) and their distribution across 9 sampling locations (inner ring). For each location, 3 samples were collected per species (i.e., 60 samples per location). The final dataset consisted of 540 morphometric observations collected from 20 freshwater fish species sampled across 9 locations of the Indus, Chenab, and Jhelum river systems. For each species, three specimens were recorded from each sampling location, resulting in a balanced dataset structure (20 species × 9 locations × 3 specimens = 540 observations). The dataset was stored in CSV format containing 22 feature columns and one header row, giving a total spreadsheet size of 541 rows.

Table 1. Scientific and common names of selected fish species with sample size data set

Sr. No.	Species (Scientific Name)	Local Name	Sample Size
1	*Catla catla*	Thaila	27
2	*Labeo calbasu*	Kalbans	27
3	*Labeo gonius*	Ghonya –Goonia	27
4	*Labeo bata*	Bata	27
5	*Labeo rohita*	Rahu	27
6	*Cyprinus carpio*	Gulfam	27
7	*Cirrhinus mrigala*	Mori	27
8	*Ompok bimaculatus*	Butter catfish, pabda	27
9	*Cirrhinus reba*	Reba	27
10	*Salmostoma phulo*	Finescaled , Razorbelly	27
11	*Systomus sarana*	Olive barb	27
12	*Heteropneustes fossilis*	Singhari – Stinging catfish	27
13	*Eutropiichthys vacha*	Vacha – Batchwa catfish	27
14	*Notopterus notopterus*	Featherback – Bronze	27
15	*Chitala chitala*	Featherback – Indian	27
16	*Bagarius bagarius*	Goonch Catfish	27
17	*Clupisoma garua*	Garua Bacha Catfish	27
18	*Wallago attu*	Malee	27
19	*Sperata seenghala*	Singhara	27
20	*Rita rita*	Rita catfish	27

Table 2. Rivers, sampling locations and coordinates of the study areas in the Punjab, Pakistan

Sr. No.	River	Location	Coordinates
1	Indus	(Mianwali (MW)	32°35′7″N 71°32′37″E
		Kallur kot (KK)	32°9'23" N and 71°15'34" E
		Dera Ghazi Ghat (DGG)	30°03′00″N 70°37′58.8″E
		Chashma Barrage (CB)	32°24′N 71°21′E
		Head Taunsa (HT)	30°54′N 70°58′E
		Head Muhammad wala (HMW)	30.2992° N, 71.3460° E
2	Chenab	Trimmu Barrage (TB)	31.1448° N, 72.1464° E
		Chiniot Pull (CP)	31°45′10″N 72°56′53″E
3	Jhelum	Rasul Barrage (RB)	32°40′49″N 73°31′15″E

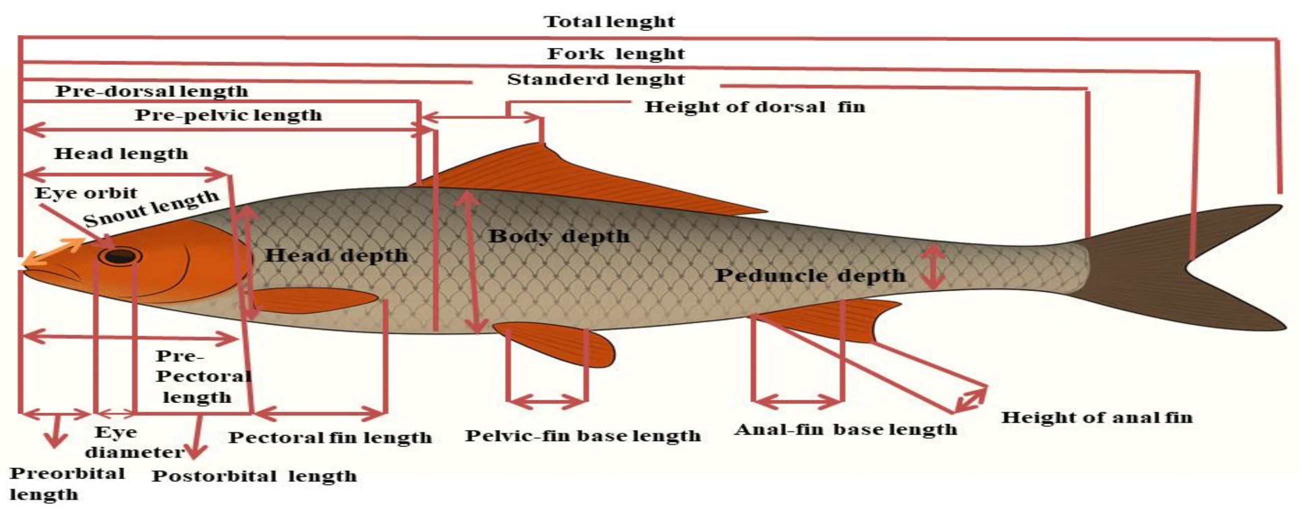

Figure 1: Morphological feature extraction of fish.

Table 3. Summary of fish morphological features used in the study

Sr. No.	Morphological Features	Short Form	Sr. No.	Morphological Features	Short Form
1	Total length	TL	**13**	Height of dorsal fin	HDF
2	Fork length	FL	**14**	Height of anal fin	HAF
3	Standard length	SL	**15**	Pectoral fin length	PFL
4	Pre-dorsal length	PrDL	**16**	Pelvic-fin base length	PBL
5	Pre-pelvic length	PPL	**17**	Anal-fin base length	ABL
6	Head length	HL	**18**	Eye orbit	EO
7	Pre-pectoral length	PPecL	**19**	Eye diameter	ED
8	Preorbital length	POL	**20**	Snout length	SnL
9	Postorbital length	PostOL	**21**	Interorbital distance	IO
10	Body depth	BD	**22**	Upper jaw length	UJL
11	Head depth	HD	**23**	Lower jaw length	LJL
12	Peduncle depth	PD	**24**	Post-dorsal length	PsDL
			25	Pre-anal length	PAL

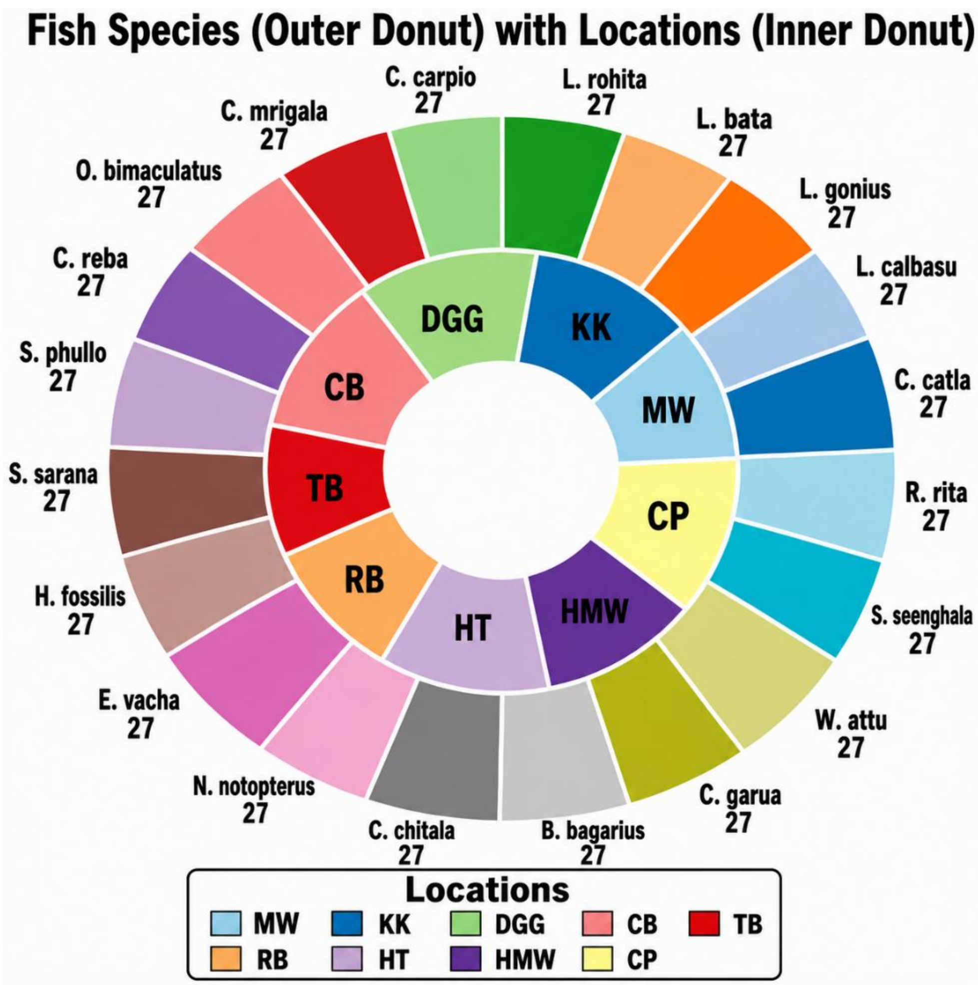

Figure 2: Schematic representation of fish species and locations.

Development and evaluation of classification framework: The classification framework was developed and evaluated using the cloud-based programming environment. In particular, the Google Colab was preferred as a primary predictive model, because it supports Python-based libraries and GPU (Graphics processing unit) acceleration. The integrated growth devices of the Colab were used for all codes, and the devices have been copy-offer qualified and optimized for computational efficiency. The species prediction of fish from morphological data is presented in Figure 3.

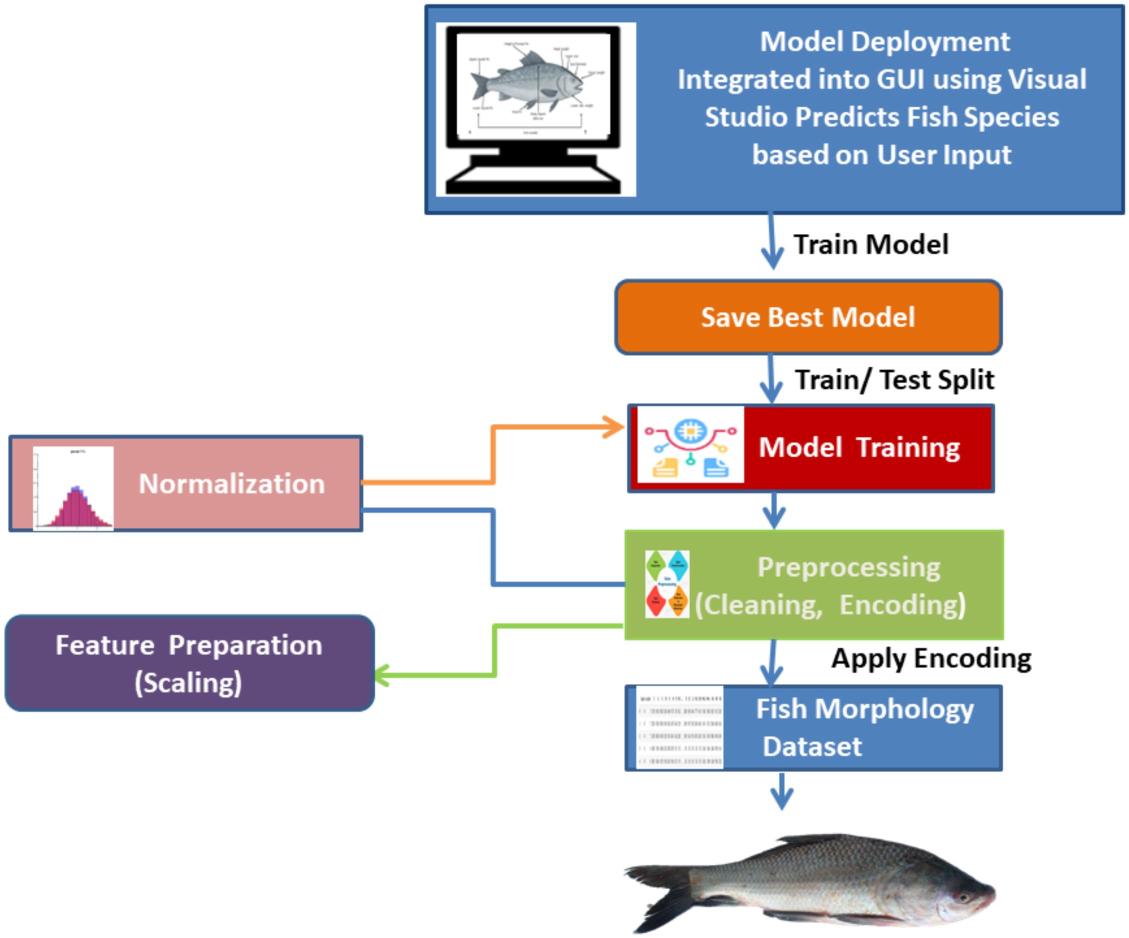

Figure 3: Schematic diagram for fish species prediction using morphological data. The sequence begins with the fish morphology dataset, followed by preprocessing steps including cleaning, encoding, feature preparation (scaling), and normalization. The processed data undergoes a train/test split for model training. The best-performing model is saved and deployed in a Visual Studio–based GUI, allowing users to predict fish species from input features.

Data Preprocessing and Feature Engineering: To ensure consistency and quality, the morphological dataset of fish species was first subjected to preprocessing steps, which include data cleaning, data imputation, and data encoding. This involved data cleansing as well as scaling numerical features into a common range through normalization. Encoder was used to convert categorical attributes to numeric values to make the model compatible. The preparation of features ensured that all the morphological measurements and categorical variables were suitable for input to the classification models.

Model Development and Application Deployment: The dataset was divided into training and testing subsets using a stratified sampling strategy to maintain proportional representation of all fish species classes. The model development and evaluation was done by an 80:20 train–test split. To minimize the risk of data leakage, pre-processing and model fitting were conducted on a training set and the testing set was then used to evaluate the model's performance. The processed morphological data was used to classify the fish species using several machine learning algorithms. The accuracy of the models was tested with Random Forest, K-Nearest Neighbors, Decision Tree and Support Vector Machine. The model with the highest performance was the Random Forest model which was saved for deployment. While some morphometric variables showed biological relationship (e.g. TL, FL, SL), all of these were kept for maintaining traditional taxonomic data as is required in ichthyological research. Random Forest was chosen due to its effectiveness in dealing with correlated variables and nonlinear feature interactions. The relative contribution of each morphometric parameter to species classification was then assessed by feature importance analysis. All code was developed and executed using Python on Google Colab, an online programming platform. The final prediction framework was deployed through a Streamlit-based graphical interface, developed in Visual Studio, to provide a user-friendly tool for prototype framework for fish species prediction. Model benchmarking was initially performed to compare the classification performance of Random Forest, K-Nearest Neighbors, Decision Tree, and Support Vector Machine algorithms using the training dataset generated from the stratified train–test split procedure. Final model validation was subsequently conducted using previously unseen testing data and stratified cross-validation analysis (10-fold stratified cross validation and receiver operating characteristics curves) to assess generalization performance and reduce evaluation bias.

Streamlit: Streamlit is an open-source framework that help to build and share elegant machine learning web tools quickly. Python library made especially for people who work with machine learning. Most data scientists know Streamlit already, so it makes their work easier. The preprocessed dataset and the trained Random Forest Classifier model were loaded into the Streamlit web application.

RESULTS

Consistency and compatibility of the dataset and models were improved by preprocessing procedures. The species had one categorical value (3), for consistency it has been changed to a numerical value. The location was label-encoded, where the original values 7, 6, and 5 were re-assigned as ordinal numbers. The data contained two categories: Species and Location. The rest of the parameters, namely Total Length (TL), Standard Length (SL), Fork Length (FL), Body Depth (BD), Head Length (HL), Head Depth (HD), Eye Diameter (ED), Preorbital Length (PrOL), Postorbital Length (PsOL), Snout Length (SnL), Interorbital Distance (IO), Upper Jaw Length (UJL), Lower Jaw Length (LJL), Predorsal Length (PrDL), Postdorsal Length (PSDL), Pre-pectoral length (PPecL), Pre-pelvic length (PPL), Pre-Anal length (PAL), Height of Dorsal Fin (HDF) and Height of Anal Fin (HAF) were continuous numerical measurements (mm) of morphological characteristics of the fish (Table 4).

The Kernel Density Estimation (KDE) plot illustrates the distribution of the normalized morphometric features used in the dataset. Because the variables were normalized during preprocessing, the x-axis represents normalized feature values rather than the original measurement units, while the y-axis represents the estimated probability density of observations. This visualization was used to assess the consistency and distributional characteristics of the input variables prior to model development (Figure 4).

Figure 5 shows the important scores of different fish morphological features using the Random Forest model. Importance of characteristics is a measure of how well each characteristic helps in the classification of fish species. Lower Jaw Length (LJL) and Upper Jaw Length (UJL) were the most important features accounting for the highest importance scores, thus these measurements were the most important in the distinction of fish species. PsDL (Post-dorsal Length), IO (Interorbital Distance), and PrOL (Preorbital Length) were other significant factors, also significantly affecting classification accuracy. The features those were relatively less important such as PsOL, TL and SL had relatively low importance value indicating lesser contribution to the species discrimination. Overall, the results showed that the most informative features for accurate fish species classification were jaw related and head related morphological measurements.

The classification report summarizes the performance of the fish species prediction model across 20 species (Figure 6A), where accuracy represents the proportion of correctly classified samples, whereas precision, recall, and F1-score provide class-specific evaluation of model performance. It shows precision, recall and F1-score values for each class indicating the model's accuracy in the identification of individual species. The vast majority achieved full marks (1.00) with consistent and accurate classification. There was, however, a slight overlap in morphological features that render some species difficult to distinguish, as evidenced by the lower performances of a few species, *C. garua*, *E. vacha*, and *S. seenghala*. In general, the model was found to be very accurate, with an accuracy of 92.59%, macro-average F1-score of 0.86 and weighted-average F1 score of 0.92, which indicates that the model is reliable in predicting the data in the dataset. In this study, the 95% confidence interval (CI) for precision (0.8804–0.9755), recall (0.8851–0.9722), and F1-score (0.8806–0.9685) provided reliable and stable estimates of performance.

The confusion matrix offers a visual comparison of actual species labels (on the vertical axis) against predicted labels (on the horizontal axis). The diagonal blocks (shaded darker blue) correspond to correct classifications, and the off-diagonal blocks (shaded lighter blue) correspond to misclassifications. Most of the predictions are along the diagonal, indicating a high classification accuracy. There are some misclassifications seen between species with close morphological similarity such as between *L. rohita* and *L. calbasu* and between *S. seenghala* and *W. attu*. The overlaps indicate the difficulties in the species differentiation of subtle anatomical differences.

Table 4. Morphological features and prediction metrics applied in fish species identification with preprocess and simple label encoding

Index	S	L	TL	SL	FL	BD	HL	HD	ED	PrOL	IO
0	3	7	211.00	161.00	185.00	52.06	49.68	47.70	7.65	18.22	28.58
1	3	7	221.00	174.00	193.00	59.54	53.62	46.79	8.44	21.39	29.49
2	3	6	212.00	165.00	181.00	52.10	51.82	46.01	8.42	20.58	29.78
3	3	5	104.53	78.06	88.08	20.12	29.30	20.97	5.15	5.15	5.15
4	3	5	100.01	85.04	92.10	19.65	26.07	19.65	5.15	5.15	5.15

Continued--

Index	S	L	UJL	LJL	PrDL	PsDL	PPecL	PPL	PAL	HDF	HAF
0	3	7	15.88	14.44	77.77	41.93	55.00	88.58	134.37	41.21	32.00
1	3	7	15.07	15.35	82.49	43.46	54.14	92.13	136.94	45.89	36.63
2	3	6	16.42	15.76	80.15	42.93	52.62	93.69	138.54	41.98	33.96
3	3	5	5.15	5.15	5.15	5.15	26.45	39.64	58.65	19.99	16.16
4	3	5	5.15	5.15	5.15	5.15	24.02	35.98	56.39	19.56	12.32

S= Species, L= Location, Total Length (TL), Standard Length (SL), Fork Length (FL), Body Depth (BD), Head Length (HL), Head Depth (HD), Eye Diameter (ED), Preorbital Length (PrOL), Interorbital Distance (IO), Upper Jaw Length (UJL), Lower Jaw Length (LJL), Pre-dorsal Length (PrDL), Postdorsal Length (PsDL), Pre-pectoral length (PPecL), Pre-pelvic length (PPL), Pre-Anal length (PAL), Height of Dorsal Fin (HDF) and Height of Anal Fin (HAF).

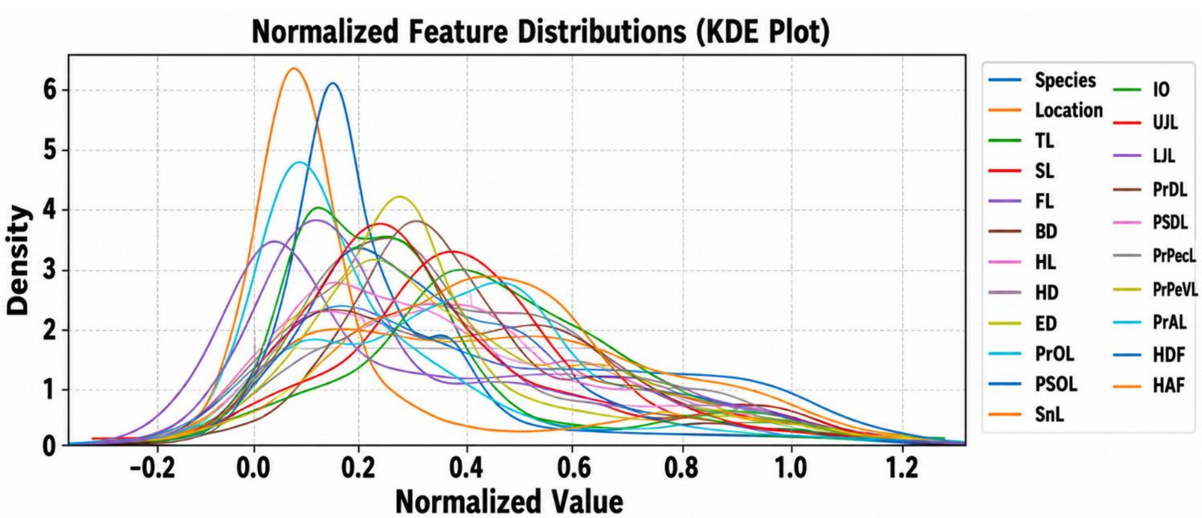

Figure 4: Kernel Density Estimation (KDE) plot showing the distribution of normalized morphometric features used in the fish species classification framework. The x-axis represents normalized feature values obtained after data preprocessing, whereas the y-axis represents kernel density estimates describing the relative distribution of observations for each feature.

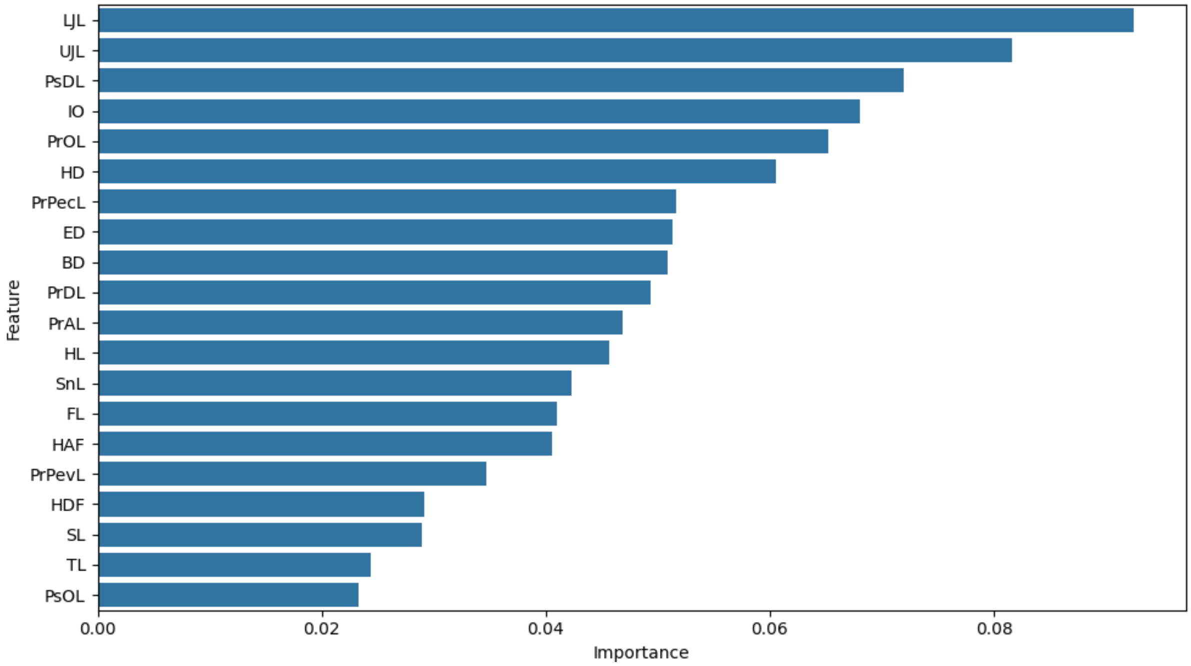

Figure 5: Feature importance analysis of morphological characteristics using the random forest classifier.

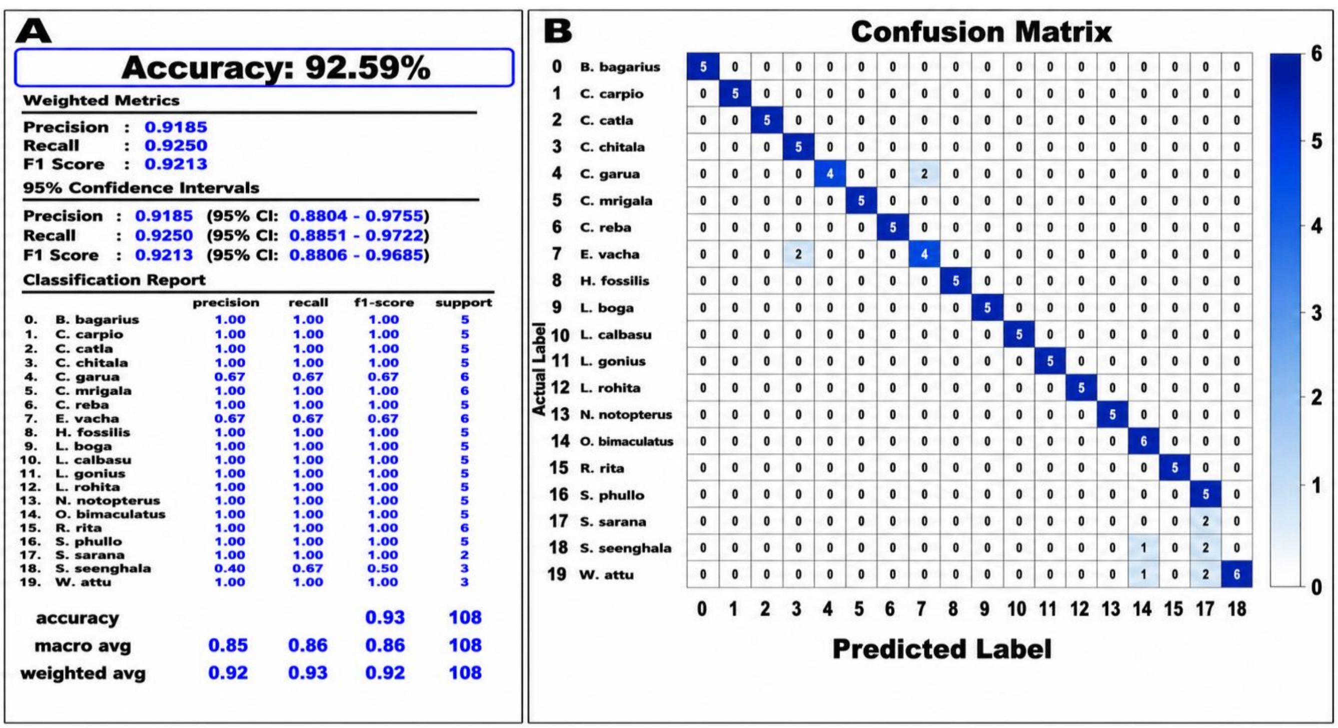

Figure 6: Performance evaluation of the fish species classification model. (A) Summary of classification performance, including overall accuracy, weighted precision, recall, F1-score, 95% confidence intervals, and class-wise metrics. (B) Confusion matrix showing the distribution of actual and predicted fish species classes, where the majority of samples are correctly classified along the diagonal.

A prototype interface was developed to demonstrate the practical applicability of the trained classification model for fish species prediction using morphometric characteristics (Figure 7). The interface was intended solely as a proof-of-concept implementation and was not subjected to formal user evaluation, expert validation, or field deployment.

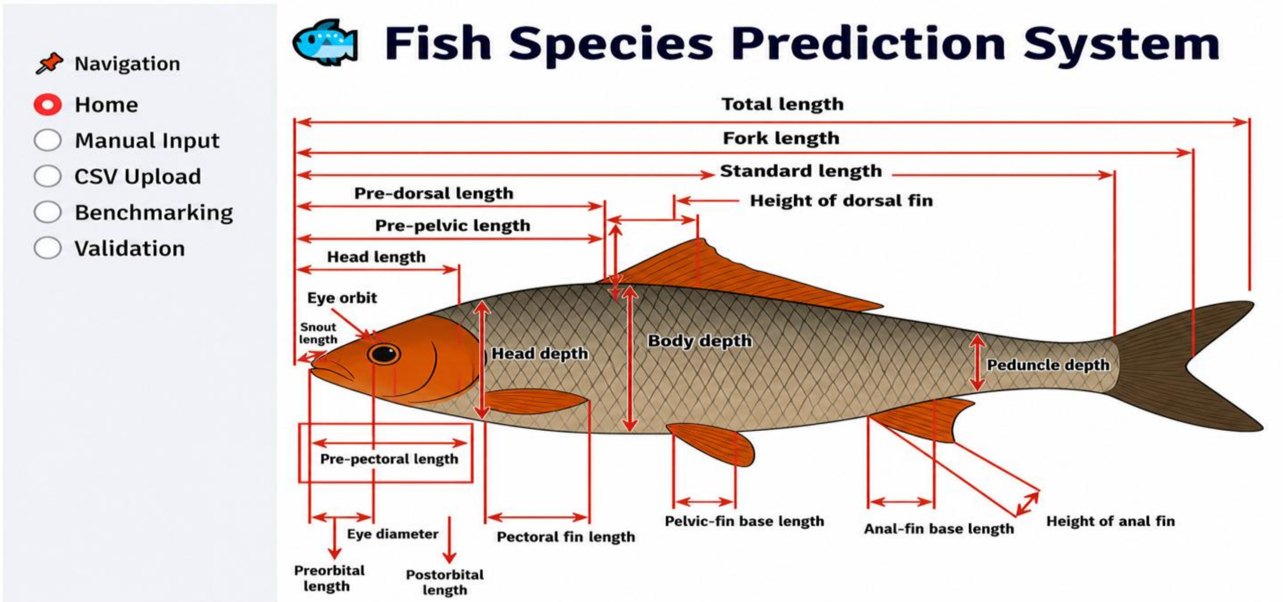

Figure 7: Prototype classification interface, a tool designed to analyze fish morphology for species identification. It features a diagram of a fish with labeled measurements and a navigation menu for user interaction.

On the fish morphological data, we compared the four models to their performance and then checked the accuracy for each model from an independent test data set as presented below.

- Random Forest was found to be the most accurate (0.93) with high accuracy and reliability in the dataset.
- Decision Tree was found with an accuracy of 0.90, which was slightly less robust than Random Forest.
- K-Nearest Neighbors (KNN) achieved an accuracy of 0.69, which was not particularly high, especially when species have similar morphological characteristics.
- Support Vector Machine (SVM), with the lowest accuracy (0.62), did not seem to be suitable for this dataset.

The results clearly showed that the ensemble-based models, particularly Random Forest, are superior to the other models and, performs better, especially for fish species classification tasks. The training and test partitions obtained from the split of the dataset were then used to benchmark the predictive performance of several machine learning algorithms: Random Forest, K-Nearest Neighbors, Decision Tree and Support Vector Machine. The independent testing subset from the stratified train–test split and further cross-validation procedures were then used for the validation analysis. The same samples were not used at the same time for the model training and final evaluation.

The performance evaluation of the proposed model is shown in figure 8 using 10-Fold Stratified Cross-Validation. The data set was divided into ten folds and each fold used 9/10 of the data set for training and the remaining 1/10 for testing. The model successfully obtained the accuracy of the 90.7% to 98.1% range in different folds. The red dashed line indicates an average accuracy of 94.63%. The low standard deviation of 2.55% shows the consistency and stability of the model across all folds.

The receiver operating characteristics (ROC) curve shows the effectiveness of the proposed machine learning model in fish species classification (Figure 9). The True Positive Rate (Sensitivity) is a useful metric to use for comparing classification models using ROC analysis. In this figure, almost all the class curves are in the upper left corner, suggesting a good classification performance. The overall Area under the Curve (AUC) of 99.89% indicates that the model is able to distinguish the fish species very well. The AUC value of 1.00 was obtained for species like *B. bagarius*, *C. catla*, *L. rohita* etc. which showed almost perfect prediction capability. Slightly lower AUC values (0.99) are only observed in a few classes, which still correspond to an excellent performance. The dotted black line in the figure is a random baseline classifier; all the ROC curves are significantly higher, showing that the proposed model has a high predictive accuracy for real-time intelligent fish classification systems with high reliability and robustness.

Overall, the validation results demonstrated that the classification model is effective in predicting fish species from morphological features, with only minor errors observed in a few species groups.

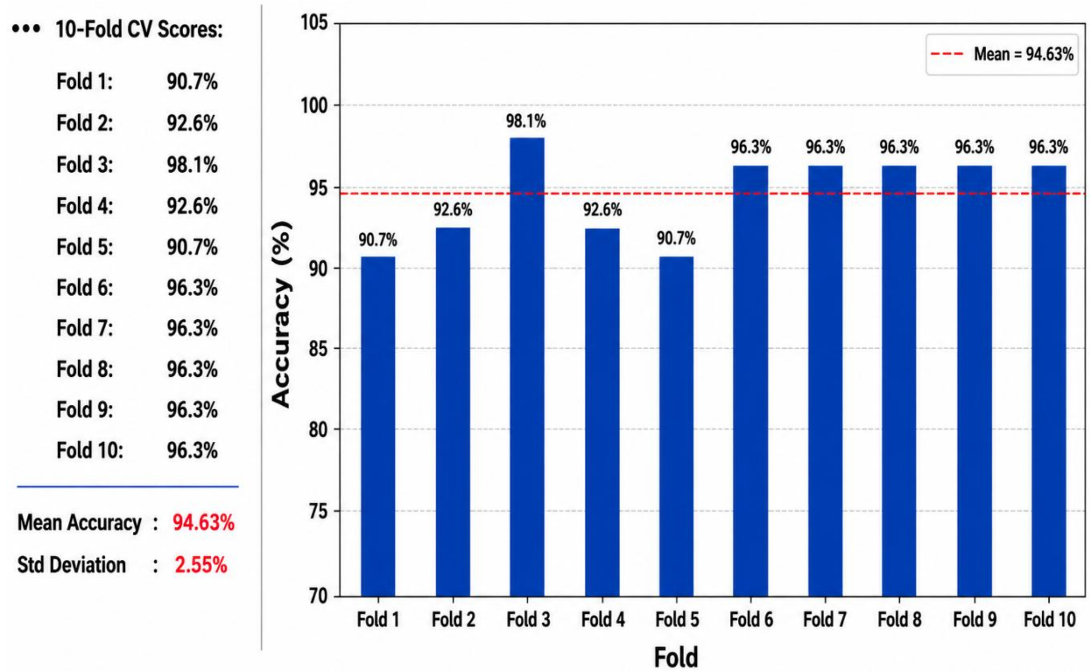

Figure 8: 10-Fold Stratified Cross-Validation analysis showing fold wise accuracy and mean accuracy (94.63%) of the proposed model.

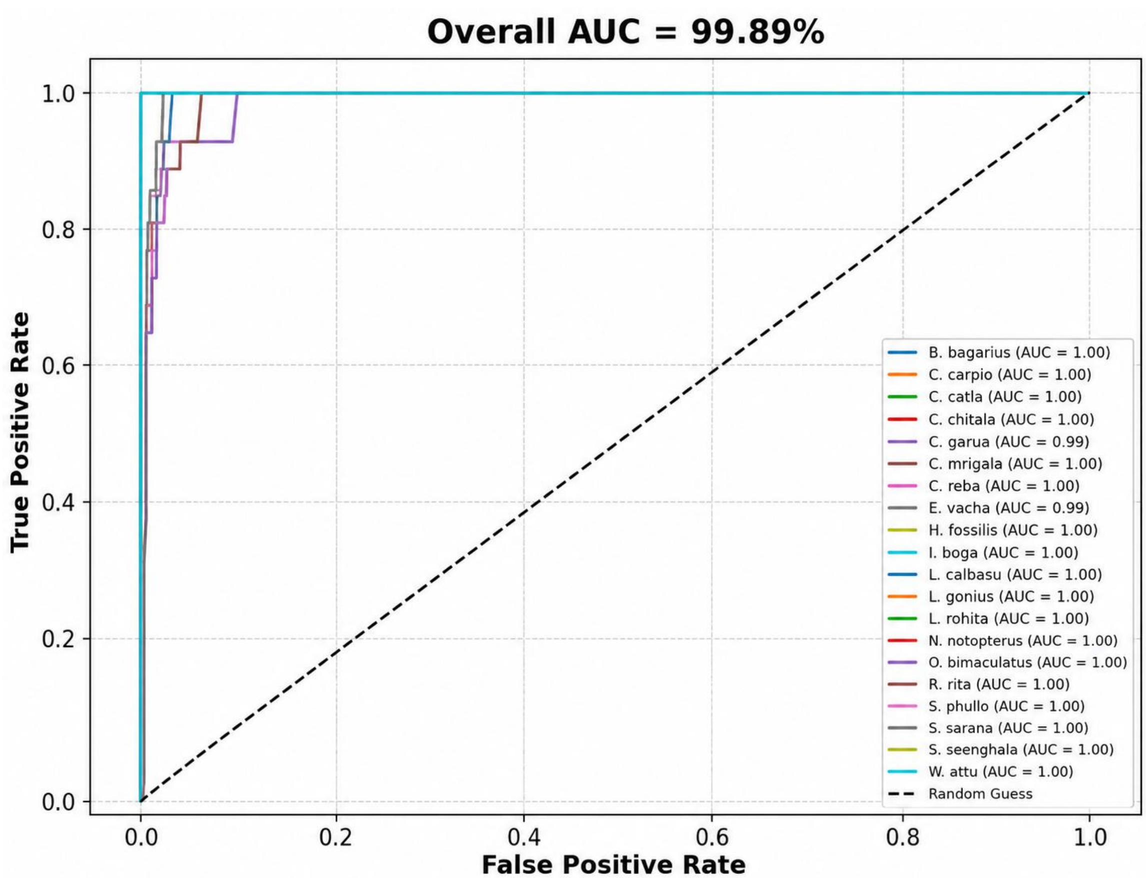

Figure 9: Receiver operating characteristics (ROC) curves for 20 fish species, showing near perfect classification performance with an overall area under the curve (AUC) of 99.89%.

DISCUSSION

This study demonstrates the applicability of integrating machine-learning techniques with morphometric analysis for freshwater fish species classification in selected river systems of Punjab, Pakistan. It shows how morphological features can be used to classify 20 freshwater fish species collected from nine locations of three different rivers within the Punjab region of Pakistan by involving three samples of each fish species collected at each site. This uniform sampling led to the creation of a database of 540 rows and 22 columns that are a reasonable starting point featuring to track biodiversity. Another row with names of the species, location of the samples and morphometrical features of the samples makes the data even more useful to ecology and taxonomical studies. Sampling every site does not seem weak or unreliable since the species appears in the majority of sites making sampling bias a significant limitation of ecological research. In the past, several imaging and machine learning methods have been investigated for fish species identification, such as multispectral imaging and deep learning frameworks (Monteiro et al. (2023) but not all of them are applicable to identifying fish species. Results of direct comparison between classification accuracy among studies are, however, challenging due to the variation in the composition of data sets, species diversity, sampling conditions, representation of features and analytical methodology. The present study therefore focuses primarily on evaluating the feasibility of morphology-based machine learning classification within the context of freshwater fish species from selected river systems of Pakistan. Likewise, Ou et al. (2025) emphasized in their scientometric review that extensive literature is available on fish identification and highlighted the importance of high-quality datasets for enhancing biodiversity. This requirement can be fulfilled by the systematized information presented below, which at least satisfies ecological assessment that is both constant and predictable and is not determined by some intricate technological procedures. The utility of morphology to study ecology is also supported by other reports. According to Rafique and Khan (2012), there were 86 Pakistan freshwater species of special interest based on endemism, economic and conservation interest. Khan et al. (2021) and Sheikh et al. (2017) also emphasized that the populations of freshwater fish in the ecosystem of the River Indus be classified to manage the fisheries. These data are supported by the existing findings, which show that even quantitative morphological data such as total length, body depth, and head sizes can be applied to provide sufficiently good status recognitions. Morphology is the other feasible substitute and cleaner replacement, rather than the other processes which necessitate the utilization of the expensive imaging processes in certain locations that are not as developed as the rest of the advanced technologies. The present framework may be particularly valuable in regions where advanced imaging systems, molecular laboratories, and automated underwater monitoring technologies remain economically inaccessible. Manual identification in the field remains a very time consuming and taxonomically demanding process, and is still widely used for fisheries management and biodiversity monitoring in many developing countries. The proposed method is a relatively low-cost and scalable method in the preliminary species identification process based on morphological characteristics and machine learning classification. These systems could also help researchers, fisheries departments and conservation authorities in standardizing data and help with fast ecological assessments in freshwater.

Feature importance analysis (Figure 5) of the Random Forest classifiers with our morphometric data showed that the most important morphological features that discriminated between the studied fish species were Lower Jaw Length (LJL) and Upper Jaw Length (UJL). Jaw morphology is closely associated with feeding behavior, prey capture strategies, trophic specialization, and ecological niche differentiation among fish species. The size of the jaw may be related to different feeding habits and food resources, thus being useful for species discrimination. Our study is supported by the study of Wainwright and Richard (1995), who studied the predicting patterns of prey use from the morphology of fishes. They revealed that when prey were categorized into functional groups, reflected the specific performance features that were important in capturing and handling them, and the differences among habitats in the available prey resources were taken into account, general patterns could be found in morphology-diet relations that crossed phylogenetic boundaries.

The 10-fold stratified cross-validation also showed the stability and consistency of the model, with an average accuracy of 94.8%. The mean cross validation accuracy was good, but there was some variability when performing fold-wise cross validation. This variation could be due to a combination of differences in sample composition, morphometric similarity of species, and natural variation in the data. Overall the model showed good performance and this was good, but more validation with larger and geographically independent datasets would be useful, to further evaluate the stability and transferability of the classification.

Lastly, this study has demonstrated that morphology-based identification is a valid and justifiable method of fish taxonomy and biodiversity surveillance. These data are sorted and equalized in a manner that illustrates that systematic sampling using appropriate morphometric features can provide cost-efficient data that can be replicated on a mass scale with a wide range of applications. In this study, the practical implementation interface was mainly developed to be a demonstration of the applicability of the model and not a fully validated operational system. Formal usability assessment, end user evaluation or deployment testing in the field was not performed in the current study. In this regard, the interface must be seen as a demonstration of a prototype system, not a validated decision support system. Another limitation concerns the evenness of the experimental data, which consisted of equal numbers of individuals for each species at each of the sampling sites. The datasets from natural fisheries are frequently unevenly distributed in species abundance, incomplete and have varying sampling conditions, which makes this design useful for decreasing class imbalance in the model training process and easier when comparing species, but not when dealing with natural fisheries datasets. Thus, the classification results presented in the current study may not be fully representative of the difficulty with which the monitored scenarios may occur in the real world. Future investigations should therefore evaluate model performance using larger, more diverse, and naturally imbalanced datasets to assess robustness under operational field conditions.

Although the classification performance that was achieved in the present study is promising, there are a number of limitations that should be taken into consideration. Data collection is still limited in capturing the full scale of intraspecific morphological variation due to age, sex, environment and seasonal effects. In addition, age class, sex, and reproductive maturity status were not recorded during specimen collection and therefore could not be evaluated as potential sources of morphometric variation. Since ontogenetic development, sexual dimorphism, and reproductive condition may influence body proportions in fishes, future studies should incorporate these biological variables to improve model generalizability and ecological interpretation. Although measurements were independently verified and cross-checked prior to analysis, formal inter-observer reliability statistics were not available and should be quantified in future investigations. Future studies should perform direct comparisons between machine-learning predictions, traditional identification keys, and expert assessments to evaluate relative strengths and limitations under field conditions.

Besides, some closely related species showed overlapping morphological characters that can lead to a controversial classification. The misclassifications observed between certain species pairs also provide useful biological insight. In certain cases, confusion between *L. rohita* and *L. calbasu* may be due to the fact that the species are similar in body proportions and morphometric characteristics, which are often used to identify the species. Likewise, misclassifications between *S. seenghala* and *W. attu* might be linked to overlapping values of some of the body and head measurements used in the data. These misclassifications are likely to occur if the species have similar morphological characteristics, especially if identification is based on quantitative morphometric features only. These results underscore the need for a multi-character diagnosis and illustrate that some closely related and/or morphologically similar species can still be difficult to distinguish solely by morphology.

Manual morphometric measurements also have small variations which could affect the performance of the model. To enhance the generalizability of the models and ecological interpretation, future studies should include detailed biological and ecological metadata and larger sample populations. Future studies should therefore incorporate larger and more geographically diverse datasets, image-based features, molecular confirmation methods, and advanced deep learning approaches to improve classification robustness and ecological generalizability. Morphometric measurements combined with computer vision methods can yield more comprehensive and scalable solutions for biodiversity monitoring in aquatic systems.

Conclusions: This study demonstrates the successful integration of traditional morphometric analysis with machine learning frameworks to classify 20 economically important freshwater fish species from the Indus, Chenab, and Jhelum river systems in Punjab, Pakistan. By leveraging 22 distinct morphological measurements from a balanced dataset of 540 specimens, the framework provides a scalable, cost-effective alternative to expensive imaging systems or molecular techniques, which are often inaccessible in resource-limited regions. Ultimately, this approach underscores the enduring value of physical morphology in modern computational taxonomy, offering a reliable blueprint for automated biodiversity monitoring and sustainable fisheries management that can be empirically scale in future work.

Data availability statement: All relevant data regarding this manuscript will be available on request from the corresponding author.

Conflict of interest: The authors declare no conflicts of interest.

REFERENCES

Abinaya, N.S., D. Susan and R.K. Sidharthan (2022). Deep learning-based segmental analysis of fish for biomass estimation in an occulted environment. Comp. Electron. Agric. 197:106985. https://doi.org/10.1016/j.compag.2022.106985

Al-Abri, S., S. Keshvari, K. Al-Rashdi, R. Al-Hmouz and H. Bourdoucen (2025). Computer vision based approaches for fish monitoring systems: a comprehensive study. Artif. Intell. Rev. 58: 185. https://doi.org/10.1007/s10462-025-11180-3

Allken, V., N.O. Handegard, S. Rosen, T. Schreyeck, T. Mahiout and K. Malde (2019). Fish species identification using a convolutional neural network trained on synthetic data. ICES J. Mar. Sci. *76*(1): 342–349. https://doi.org/10.1093/icesjms/fsy147

Chandak, A.V., P. Bante, A. Ghaiwat and N. Malani (2023). Automatic identification of fish species using machine learning. Proc. OITS Int. Conf. Inform. Technol. Raipur, India. 2023:364-368. https://doi.org/ 10.1109/OCIT59427.2023.10430793

Farooq, Z., M. Ramzan, M. Bilal, M. Attique, T. S. Chung and A. Naz (2026). A multi-class framework for fish species classification using deep learning technique. Plos One, *21*(2), e0342901. https://doi.org/10.1371/journal.pone.0342901

Ghouri, M.Z., M. Ismail, M.A. Javed, S.H. Khan, N. Munawar, A.B. Umar, S.O. Aftab, S. Amin, Z. Khan and A. Ahmad (2020). Identification of edible fish species of Pakistan through DNA barcoding. Front. Mar. Sci. 7: 554183. https://doi.org/ 10.3389/fmars.2020.554183

Jalal, A., A. Salman, A. Mian, M. Shortis and F. Shafait (2020). Fish detection and species classification in underwater environments using deep learning with temporal information. Ecol. Inform. 57: 101088. https://doi.org/10.1016/j.ecoinf.2020.101088

Khan, M.Q., M.Z. Anjum, M. Adnan, A. Khan, H. Zahid, J. Nawab, S.Z. Safi, M.I.A. Shah, A. Kamil and A. Ali (2021). Genetic Diversity of *Schizothorax*, *Tor* and *Mystus* spp. in Khyber Pakhtunkhwa, Pakistan: Species of Economic Importance. Pak. J. Zool. *53*(3): 1099-1109. https://dx.doi.org/10.17582/journal.pjz/20190906180937

Kühn, B., A. Cayetano, J.I. Fincham, H. Moustahfid, M. Sokolova, N. Trifonova, J.T. Watson, J.A. Fernandes-Salvador and L. Uusitalo (2025). Machine learning applications for fisheries—at scales from genomics to ecosystems. Rev. Fish. Sci. Aqua. *33*(2): 334-357. https://doi.org/ 10.1080/23308249.2024.2423189

Li, D. and L. Du (2022). Recent advances of deep learning algorithms for aquacultural machine vision systems with emphasis on fish. Artif. Intell. Rev. *55*(5): 4077–4116. https://doi.org/10.1007/s10462-021-10102-3

López-Barajas, S., P.J. Sanz, R. Marín-Prades, A. Gómez-Espinosa, J. González-García and J. Echagüe (2024). Inspection operations and hole detection in fish net cages through a hybrid underwater intervention system using deep learning techniques. J. Mar. Sci. Eng. *12*(1): 80. https://doi.org/10.3390/jmse12010080

Mirza, M.R. (2003) Checklist of freshwater fishes of Pakistan. *P*akistan J. Zool. (Supplement Series) 3: 1–30.

Mohammadisabet, A., R. Hasan, V. Dattana, S. Mahmood and S. Hussain (2025). CNN-based optimization for fish species classification: Tackling environmental variability, class imbalance, and real-time constraints. Information *16*(2): 154. https://doi.org/10.3390/info16020154

Monteiro, F., V. Bexiga, P. Chaves, J. Godinho, D. Henriques, P. Melo-Pinto, T. Nunes, F. Piedade, N. Pimenta, L. Sustelo and A.M. Fernandes (2023). Classification of fish species using multispectral data from a low-cost camera and machine learning. Remot. Sens. *15*(16): 3952. https://doi.org/ 10.3390/rs15163952

Mujtaba D.F. and N.R. Mahapatra (2022). Hierarchical deep learning models for identification of fish species. Proc. Int. Conf. Comp. Sci. Comp. Intl. (CSCI), Las Vegas, NV, USA. 2022: 1588-1593.

Ou, L., L. Lu, W. Qian and B. Liu (2025). Application of artificial intelligence in fish information identification: a scientometric perspective. Front. Mar. Sci. 12: 1575523. DOI: 10.3389/fmars.2025.1575523

Rafique, M. and N.U.H. Khan (2012). Distribution and status of significant freshwater fishes of Pakistan. Rec. Zool. Surv. Pak. 21: 90-95.

Rashid, M. K., I.R. Salman, A.L. Obaid, S.A.D.H. Hassan, M.R. Al-musawi and M. Al-Saady (2024). Application of machine learning in predicting sources of water pollution in the Euphrates and Tigris rivers in Iraq. Int. J. Aqua. Biol. *12*(6): 581-589. https://doi.org/10.22034/ijab.v12i6.2421

Sajjad, A., F. Jabeen, M. Ali and S. Zafar (2023). DNA barcoding and phylogenetics of *Wallago attu* using mitochondrial *COI* gene from the River Indus. J.K.S.U.S. 35:102725. https://doi.org/10.1016/j.jksus.2023.102725

Sheikh, M., M.Y. Laghari, P.K. Lashari, A.R. Khooharo and N.T. Narejo (2017). Current status of three major carps (*Labeo rohita*, *Cirrhinus mrigala* and *Catla catla*) in the downstream Indus River Sindh. Fish Aqua. J. *8*(3): 1000222. https://doi.org/ 10.4172/2150-3508.1000222

Tseng, C.H. and Y.F. Kuo (2019). Detecting and counting harvested fish and measuring fish body lengths in video using deep learning methods. Proc. ASABE Annual International Meeting. Boston Massachusetts, USA. 2019: 1900408. https://doi.org/ 10.13031/aim.201900408

Villon, S., C. Iovan, M. Mangeas and L. Vigliola (2022). Confronting deep-learning and biodiversity challenges for automatic video-monitoring of marine ecosystems. Sensors *22*(2):497. https://doi.org/10.3390/s22020497

Wainwright, P.C. and B.A. Richard (1995). Predicting patterns of prey use from morphology of fishes. Environ. Biol. Fish 44: 97–113. https://doi.org/10.1007/BF00005909

Wang, G., A. Muhammad, C. Liu, L. Du and D. Li (2021). Automatic recognition of fish behavior with a fusion of RGB and optical flow data based on deep learning. Animals *11*(10): 2774. https://doi.org/10.3390/ani11102774