Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

BirdNova: A Web-Based Bird Species Identification Platform Using Separate Image and Audio Deep Learning Models

Vedika Shantisagar Chavare[1]
Sharad institute of technology college of engineering
Nikita Shrenik Patil[1]
nikitapatil3216@gmail.com

Sharad institute of technology college of engineering
Suvidha Vilas Kamble[1]
Sharad institute of technology college of engineering
Minakshi Popat Mane[1]
Sharad institute of technology college of engineering

Research Article

Keywords: bird species identification, deep learning, image classification, audio classification, convolutional neural network, Mel-spectrogram, biodiversity monitoring, web application

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11772764/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Bird species identification supports biodiversity research and wildlife monitoring, but manual identification can be difficult when species have similar visual characteristics or vocalisations. This paper presents BirdNova, a web-based application that provides bird identification through two independent deep-learning workflows: image classification and audio classification. The image workflow resizes input photographs to 224 × 224 pixels and processes them using a convolutional neural network (CNN). The audio workflow converts recordings to mono audio at 22,050 Hz, standardises their duration to five seconds by trimming or padding, and generates Mel-spectrogram representations for a separate audio classifier. The two models are accessed through a common Flask-based interface but do not perform feature-level or decision-level fusion. The image classifier was evaluated on an independent test set of 1,753 images covering 25 bird species. It achieved 80.03% test accuracy, 81.49% weighted precision, 80.03% weighted recall, and 79.93% weighted F1-score. Validation accuracy was 95.96%, a difference of 15.93 percentage points from independent test accuracy. This gap indicates that generalisation requires further investigation. The current manuscript documents the audio workflow but does not report a complete quantitative evaluation of that module; therefore, no audio performance claim is made. BirdNova is presented as an application-level integration of separate image and audio classifiers, with its evaluation limitations explicitly identified.

1. Introduction

Birds are important components of ecosystems and are frequently used as indicators in biodiversity assessment and environmental monitoring. Reliable species identification can support ecological surveys, conservation planning, and the study of changes in species distribution. Traditional identification relies on visual observation and listening to calls or songs. These tasks may require specialist knowledge and can be difficult when observations are brief, images are unclear, species have similar appearances, or recordings contain background noise.

Deep-learning methods have advanced both image recognition and bioacoustic classification. For audio, convolutional networks operating on spectrogram representations have been used to identify large numbers of bird species and to analyse long-duration soundscapes [1–4]. For images, fine-grained recognition methods have addressed the challenge of distinguishing species with similar visual patterns through part-based detection, feature enhancement, and contrastive learning [5,6]. These studies also show that evaluation on independent data and under varied environmental conditions is important when assessing practical reliability.

This paper describes BirdNova, a web-based application that makes two independent classification workflows available through a common interface. A user can submit either a bird photograph or an audio recording. The selected input is sent to its corresponding model; the image and audio models are not combined and the system does not currently perform audiovisual feature fusion. This distinction is

important because the presence of two input options in one application does not, by itself, constitute a multimodal fusion model.

The work is implementation-oriented. Its focus is the integration of separate image and audio inference pipelines, their input-preprocessing procedures, and the presentation of predicted species, confidence information, and descriptive bird information through a web application. The image classifier is quantitatively evaluated using an independent test set. The audio pipeline is described at the implementation level, but its quantitative performance is not reported in the available experimental record. Accordingly, the paper does not claim that the audio module has been experimentally validated to the same extent as the image module.

1.1 Objectives and contributions

The objectives of this work are: (i) to implement an image-based bird species classification workflow using a CNN; (ii) to implement a separate audio-based workflow using Mel-spectrogram input; (iii) to connect both workflows to a common web application; and (iv) to report the available independent-test results for the image classifier and discuss the observed validation-to-test performance gap.

The contributions are limited to the implementation and evaluation evidence available in the current project: (a) a common web interface for two independent bird-identification pipelines; (b) a description of the image and audio preprocessing and inference workflows; and (c) an independent image-classification evaluation across 25 species, including weighted and macro-averaged metrics and selected species-level results. No new fusion algorithm, novel CNN architecture, or measured improvement from combining modalities is claimed.

BirdNova is an implementation-oriented web application that provides two independent bird-species classification workflows through a common interface. A user submits either an image or an audio recording, and the application routes the input to the corresponding classifier. The image workflow uses a CNN-based model, while the audio workflow uses a separate model operating on Mel-spectrogram input. The models do not share features, combine prediction scores, or perform joint audiovisual inference. The contribution reported here is the implementation of the two separate inference paths within a Flask application, the presentation of predictions and associated bird information, and the quantitative evaluation of the image classifier on the reported independent test set. The work does not propose a new fusion algorithm or claim that combining image and audio inputs improves classification performance.

2. Related work

Research on automatic bird identification spans acoustic recognition, soundscape analysis, and fine-grained image classification. Stastny et al. [1] used acoustic features including human factor cepstral coefficients and a hidden Markov model-based approach to classify vocalisations from 18 bird species. Kahl et al. [2] described the BirdCLEF 2021 soundscape challenge, highlighting the difficulty of identifying

bird calls in continuous recordings and varied ecological conditions. BirdNET, presented by Kahl et al. [3], demonstrated deep-learning-based identification across hundreds of species and evaluated performance on both focal recordings and soundscapes. Gupta et al. [4] compared recurrent convolutional neural networks for large-scale bird sound classification and considered overlapping vocalisations and background noise. Liu et al. [8] investigated ensemble multi-scale CNNs for birdsong classification, while Xie et al. [9] reviewed deep-learning approaches to bird-vocalisation recognition and identified dataset diversity and generalisation as continuing challenges.

Image-based approaches address the fine-grained visual differences between species. Yi et al. [5] proposed a part-based bird-recognition method using an improved YOLOv5 architecture. Yang et al. [6] combined multi-scale feature extraction, feature enhancement, and contrastive learning for bird image recognition. These studies use dedicated model designs and quantitative experiments to assess recognition performance.

Research has also begun to investigate audiovisual fusion. Xie et al. [7] proposed a multi-stage audiovisual fusion network for fine-grained bird species classification, using dedicated audio and visual feature extractors and a fusion objective. This is materially different from the current BirdNova implementation: BirdNova exposes separate image and audio models through one application and does not combine their representations. The distinction is retained throughout this paper to avoid overstating the implemented functionality.

The present work therefore addresses an application-integration problem rather than proposing a new classification architecture. The practical value and generalisability of this integration need to be assessed against the reproducibility information and additional experiments identified in Sections 4–7.

3. BirdNova system overview

BirdNova is a web-based bird species identification application. It accepts one input modality per prediction request: an image or an audio recording. The application routes the uploaded file to the corresponding processing pipeline. The image pipeline resizes the photograph to 224 × 224 pixels and passes it to the image CNN. The audio pipeline converts the recording to mono, resamples it to 22,050 Hz, standardises it to five seconds, and computes a Mel-spectrogram for the audio classifier. Each classifier independently returns a species prediction and a model confidence value. The application then presents the prediction and retrieves descriptive information for the predicted species.

The application uses a Flask backend to connect the user interface with the trained models and the bird-information retrieval component. The models are maintained separately, allowing the image and audio components to be updated independently. The current implementation does not synchronise image and audio inputs, align features, combine prediction scores, or generate a joint audiovisual prediction.

The architecture figure has been revised to avoid implying that ImageNet or UrbanSound8K was used for pre-training. Dataset source information is described in Section 4.1.

3.1 Image classification workflow

When an image is uploaded, the application receives the file and resizes it to 224 × 224 pixels. The preprocessed image is converted to the numerical input format expected by the trained CNN. The model produces class scores for the supported bird categories. The class with the highest score is returned as the predicted species, together with the confidence value exposed by the application. The exact network architecture, training configuration, class mapping, and any normalisation operations must be documented to make this workflow reproducible.

3.2 Audio classification workflow

When an audio file is uploaded, the application converts the signal to mono and resamples it to 22,050 Hz. Recordings longer than five seconds are trimmed; shorter recordings are padded to the target duration. A Mel-spectrogram is then generated and supplied to the separate audio classification model. The model returns a predicted class and confidence value. The present manuscript does not report an independently measured audio test-set size or audio classification metrics. Consequently, this section describes the implemented processing path only and should not be interpreted as evidence of measured audio accuracy.

3.3 Web application and output

The Flask application manages file upload, input routing, model inference, and presentation of results. After inference, the interface displays the predicted species and the confidence value returned by the selected classifier. A bird-information retrieval component provides descriptive information associated with the predicted species. The image and audio requests are processed independently; the application does not combine outputs from both models.

4. Materials and methods

4.1 Datasets and class composition

The image dataset used for the image-classification component was the Indian Birds dataset obtained from Kaggle under the dataset identifier ichhadhari/indian-birds [10]. The local dataset structure inspected by the project team contains separate training and test directories, each organised into 25 species-specific class folders. The folder structure confirms the 25-class organisation, but the exact number of images used in each model run, per-class counts, and the relationship between the downloaded test directory and the 1,753-image evaluation set could not be established from the records available for this manuscript. Accordingly, the 1,753 images are reported only as the test-set size recorded in the original evaluation results; they are not assumed to represent the entire Kaggle test directory. The dataset version, download date, and licence terms applicable to the downloaded copy were not retained in the available project records. The Kaggle listing should be consulted for the terms applicable to the version used.

The audio component uses bird-vocalisation recordings as input to a separate classification workflow. The project records available for this manuscript do not identify the audio dataset source, version, licence, number of recordings, duration distribution, class-wise sample counts, or data-splitting procedure. For this reason, the audio component is described as an implemented inference workflow only. No claim is made that the audio data constitute a documented or independently benchmarked evaluation dataset.

4.2 Image preprocessing

For image inference, each uploaded image is resized to 224 × 224 pixels to match the input dimensions expected by the image classifier. The available implementation description does not provide a verified specification for channel ordering, pixel scaling, interpolation method, or training-time augmentation. These operations are therefore not further specified here, and no particular normalisation or augmentation procedure is claimed.

4.3 Audio preprocessing

Audio recordings are converted to mono and resampled to 22,050 Hz. Recordings longer than five seconds are trimmed, while shorter recordings are padded to the target duration. A Mel-spectrogram is then generated and passed to the separate audio classification model. The exact Mel-spectrogram configuration—including the number of Mel bands, FFT size, hop length, window function, frequency limits, logarithmic scaling, and normalisation—was not retained in the project records available for this manuscript. These parameters are therefore not stated or inferred.

4.4 Model configuration and training

BirdNova uses separate image and audio classification models. The image model is a CNN-based classifier with an input size of 224 × 224 pixels. The audio model uses Mel-spectrogram input. The trained model files are integrated into the Flask application as separate model assets. The exact layer-by-layer architectures, parameter counts, initialisation methods, loss functions, optimisers, learning rates, batch sizes, epoch counts, random seeds, hardware configurations, and software versions are not documented in the project records available for this manuscript. Consequently, no unverified training configuration is reported.

The audio classifier is maintained separately from the image classifier and is loaded by the application for audio inference. The available project records do not contain a sufficiently detailed training log or verified provenance statement to reproduce its training procedure. The manuscript therefore limits its description to the implemented preprocessing and inference workflow and does not attribute unverified pre-training sources or report audio training hyperparameters.

4.5 Evaluation metrics

The image-classification results are reported using accuracy, precision, recall, and F1-score. Precision, recall, and F1-score are presented as both weighted averages, which account for class support, and

macro averages, which assign equal weight to each class. The reported evaluation is a single-prediction classification task across 25 bird-species categories. The numerical results below are transcribed from the project evaluation results supplied for this manuscript. The original prediction file, class-wise support counts, confusion matrix, and exact metric-library configuration were not available for independent recalculation during manuscript preparation; accordingly, the reported values should be interpreted as the recorded project results.

5. Experimental results and discussion

5.1 Overall image-classification performance

The image classification model was evaluated on an independent test set of 1,753 images from 25 bird species. The reported independent test accuracy was 80.03%. Weighted precision, recall, and F1-score were 81.49%, 80.03%, and 79.93%, respectively. Macro precision, recall, and F1-score were 81.10%, 79.43%, and 79.38%, respectively. Validation accuracy was 95.96%. The results are summarised in Table 1.

Table 1
Overall image-classification results reported for the independent test set.

Evaluation measure	Reported value
Number of independent test images	1,753
Number of species	25
Validation accuracy	95.96%
Independent test accuracy	80.03%
Weighted precision	81.49%
Weighted recall	80.03%
Weighted F1-score	79.93%
Macro precision	81.10%
Macro recall	79.43%
Macro F1-score	79.38%

5.2 Species-level performance

Performance varied across the species for which class-level metrics are available. Asian-Green-Bee-Eater had the highest F1-score among the species listed in Table 2 (97.01%), while Forest-Wagtail had the lowest listed F1-score (60.00%). Common-Myna and Northern-Lapwing also had comparatively lower F1-scores. These differences indicate that aggregate metrics alone do not describe performance

uniformly across all classes. The reasons for class-level variation cannot be established from the reported metrics alone; they require examination of class support, image quality, and misclassified examples.

Table 2
Selected species-level image-classification metrics.

Bird species	Precision	Recall	F1-score
Asian-Green-Bee-Eater	98.48%	95.59%	97.01%
Ruddy-Shelduck	95.92%	94.00%	94.95%
Common-Kingfisher	92.50%	93.67%	93.08%
Hoopoe	93.10%	92.50%	92.57%
Indian-Pitta	92.54%	92.54%	92.54%
Forest-Wagtail	67.35%	54.10%	60.00%
Common-Myna	71.93%	66.13%	68.91%
Northern-Lapwing	71.74%	66.00%	68.75%
Red-Wattled-Lapwing	71.43%	68.49%	69.93%
Common-Tailorbird	80.43%	62.71%	70.48%

5.3 Validation-to-test performance difference

The validation accuracy of 95.96% exceeded the independent test accuracy of 80.03% by 15.93 percentage points. This difference is an important finding and should not be described as evidence of strong generalisation. It indicates that performance on the validation data did not transfer to the independent test set to the same extent. Potential explanations include differences in data distribution, class composition, image backgrounds, image quality, or overlap between data sources; however, the available results do not identify the cause. Further analysis should compare the validation and test splits, inspect duplicate images, report per-class support, and examine the confusion matrix and representative errors.

Table 3
Difference between validation and independent test accuracy.

Measure	Value
Validation accuracy	95.96%
Independent test accuracy	80.03%
Absolute difference	15.93 percentage points

5.4 Audio module evaluation status

The current project description specifies the audio preprocessing and inference workflow, but the available manuscript does not report the size or composition of an independent audio test set, quantitative audio metrics, or a confusion matrix. It is therefore not possible to compare audio performance with image performance or to claim that the audio classifier is accurate, robust, or superior to existing systems. A separate audio evaluation should be conducted using a clearly documented held-out test set before any quantitative performance statement is added.

No quantitative audio evaluation is reported in this study. The available project records do not provide a documented held-out audio test-set size, species distribution, accuracy, precision, recall, F1-score, or confusion matrix. The audio component is therefore presented as an implemented inference workflow, and no audio performance estimate or comparison with the image classifier is made.

5.5 Discussion

The image results show that the model achieved 80.03% accuracy on the reported independent test set, with macro F1-score of 79.38%. The species-level metrics show substantial variation, and the 15.93-percentage-point difference between validation and independent test accuracy raises questions about generalisation. These findings support a cautious interpretation: the application demonstrates a working image-classification pipeline, but its performance should be improved and more fully characterised before use in ecological decision-making. The audio module is currently described as an implemented workflow without a corresponding quantitative evaluation in this paper.

Compared with research systems that report dedicated architectures, larger datasets, independent benchmarks, and detailed experiments [3–7], the present work has a more limited application-integration contribution. To strengthen its technical value for a multimedia systems journal, the authors should add reproducibility details, a system-level evaluation (for example, inference latency, file-size limits, failure handling, and resource use), and an audio evaluation if the data are available. Any comparison with other methods must use comparable datasets and evaluation protocols; results from different datasets should not be presented as a direct performance ranking.

6. Limitations and threats to validity

This study has several limitations. First, the validation accuracy is substantially higher than the independent test accuracy, and the cause of this gap has not yet been established. Second, the manuscript does not currently provide the exact dataset sources, class-wise sample counts, or complete split protocol needed to reproduce the image evaluation. Third, the audio pipeline is described without quantitative results on a held-out audio test set. Fourth, the manuscript does not report inference latency, memory use, throughput, or usability evaluation of the web application. Fifth, the current application does not fuse image and audio features or predictions; users submit one modality at a time. Finally, the reported class-level table contains selected species rather than all 25 classes, so it should not be interpreted as a complete class-wise evaluation.

These limitations should be addressed through dataset documentation, reproducible training and evaluation details, independent audio testing, error analysis, and system-level benchmarking. The application should be treated as a research prototype and not as a substitute for expert ecological identification.

7. Conclusion

This paper presented BirdNova, a web-based bird species identification application that provides separate image-based and audio-based deep-learning workflows through a common interface. The image pipeline resizes photographs to 224 × 224 pixels and uses a CNN for species classification. The audio pipeline converts recordings to mono at 22,050 Hz, standardises them to five seconds, and uses Mel-spectrogram input for a separate classifier. The two models operate independently and no audiovisual feature or decision fusion is implemented.

The image classifier was evaluated on 1,753 independent test images from 25 species and achieved 80.03% accuracy, with a weighted F1-score of 79.93% and macro F1-score of 79.38%. The 15.93-percentage-point difference between validation and independent test accuracy indicates that generalisation requires further investigation. The available manuscript does not contain a quantitative audio evaluation, so no audio performance conclusion is drawn. Future work should improve the image model's generalisation, document the datasets and training protocol, evaluate the audio classifier independently, and benchmark the web application's operational performance.

8. Future work

Future work should first address the validation-to-test performance gap through dataset auditing, duplicate detection, class-balance analysis, error inspection, and evaluation on additional independent data. The audio model should be evaluated using a documented test set and class-wise metrics. The web application can also be assessed for inference latency, memory consumption, supported file formats, error handling, and usability. A multimodal fusion architecture may be investigated as a separate future study, but it should be described as future work until a fusion model is implemented and evaluated. Additional extensions may include field testing under varied lighting and acoustic conditions, lightweight deployment, and integration of reliable geographical observation records.

Declarations

Funding:

No funding was received for this manuscript.

Competing interests:

The authors declare that they have no competing financial or non-financial interests related to this work.

Author Contribution

V.S.C. and N.S.P. contributed to the project conceptualization and system development. S.V.K. and M.P.M. contributed to data collection, testing, and evaluation. V.S.C. prepared the manuscript, including the methodology and results. N.S.P. reviewed and edited the manuscript. All authors reviewed and approved the final manuscript.

Data Availability

The datasets used in this study are publicly available from their respective sources. The processed data and evaluation results supporting the findings are available from the corresponding author upon reasonable request.

References

1. 1. J. Stastny, M. Munk, and L. Juranek, “Automatic bird species recognition based on birds vocalization,” Journal of Audio, Speech, and Music Processing, vol. 2018, article 19, 2018. https://doi.org/10.1186/s13636-018-0143-7.
2. 2. S. Kahl, T. Denton, H. Klinck, H. Glotin, H. Goëau, W.-P. Vellinga, R. Planqué, and A. Joly, “Overview of BirdCLEF 2021: Bird call identification in soundscape recordings,” in Proceedings of the Working Notes of CLEF 2021, CEUR Workshop Proceedings, vol. 2936, pp. 1437–1450, 2021. https://ceur-ws.org/Vol-2936/paper-123.pdf.
3. 3. S. Kahl, C. M. Wood, M. Eibl, and H. Klinck, “BirdNET: A deep learning solution for avian diversity monitoring,” Ecological Informatics, vol. 61, article 101236, 2021. https://doi.org/10.1016/j.ecoinf.2021.101236.
4. 4. G. Gupta, M. Kshirsagar, M. Zhong, S. Gholami, and J. Lavista Ferres, “Comparing recurrent convolutional neural networks for large scale bird species classification,” Scientific Reports, vol. 11, article 17085, 2021. https://doi.org/10.1038/s41598-021-96446-w.
5. 5. X. Yi, C. Qian, P. Wu, B. T. Maponde, T. Jiang, and W. Ge, “Research on fine-grained image recognition of birds based on improved YOLOv5,” Sensors, vol. 23, no. 19, article 8204, 2023. https://doi.org/10.3390/s23198204.
6. 6. F. Yang, N. Shen, and F. Xu, “Automatic bird species recognition from images with feature enhancement and contrastive learning,” Applied Sciences, vol. 14, no. 10, article 4278, 2024. https://doi.org/10.3390/app14104278.
7. 7. S. Xie, J. Xie, Y. Liu, L. Sha, Y. Tian, J. Dong, D. Liang, K. Pan, and J. Zhang, “Step-by-step to success: Multi-stage learning driven robust audiovisual fusion network for fine-grained bird species classification,” Avian Research, vol. 16, article 100280, 2025. https://doi.org/10.1016/j.avrs.2025.100280.

8. 8. J. Liu, Y. Zhang, D. Lv, J. Lu, S. Xie, J. Zi, Y. Yin, and H. Xu, “Birdsong classification based on ensemble multi-scale convolutional neural network,” Scientific Reports, vol. 12, article 8636, 2022. https://doi.org/10.1038/s41598-022-12121-8.
9. 10. ichhadhari, “Indian Birds,” Kaggle dataset. Available at: https://www.kaggle.com/datasets/ichhadhari/indian-birds. Dataset version and access date were not retained in the project records.
10. 9. J. Xie, Y. Zhong, J. Zhang, S. Liu, C. Ding, and A. Triantafyllopoulos, “A review of automatic recognition technology for bird vocalizations in the deep learning era,” Ecological Informatics, vol. 73, article 101927, 2023. https://doi.org/10.1016/j.ecoinf.2022.101927.

Figures

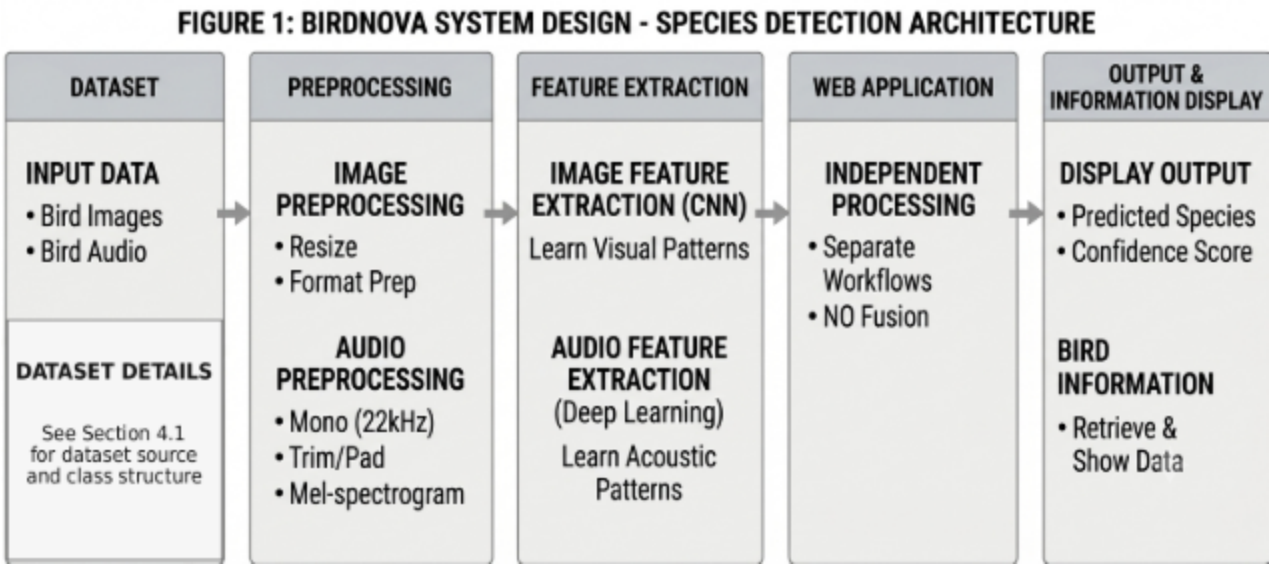


Figure 1

BirdNova system architecture showing separate image and audio processing paths connected to a common web application. The two paths do not perform feature- or decision-level fusion.