Ecological Indicators 169 (2024) 112886

Contents lists available at ScienceDirect

Ecological Indicators

journal homepage: www.elsevier.com/locate/ecolind

Original Articles

Towards transparency in AI: Explainable bird species image classification for ecological research

Samparthi V.S. Kumar, Hari Kishan Kondaveeti *

School of Computer Science & Engineering, VIT-AP University, Beside AP Secretariat, Near Vijayawada, Andhra Pradesh, India

ARTICLE INFO

Dataset link: https://www.kaggle.com/datasets/veeralakrishna/200-bird-species-with-11788-images

Keywords:
Bird species recognition
Deep learning
Explainable Artificial Intelligence (XAI)
Local Interpretable Model-Agnostic Explanations (LIME)
Transfer learning
Bird image classification

ABSTRACT

Birds are indicators of biodiversity and ecosystem health and play an essential role in maintaining the balance of natural ecosystems. However, urbanization, deforestation, and technological advancements have severely affected bird habitats, leading to a significant decline in species diversity. Manual detection and recognition of bird species based on morphological and behavioral characteristics during birdwatching and surveys are challenging and require expertise, making deep learning a promising alternative. Deep learning techniques offer the advantage of automatic identification, which can significantly enhance the efficiency and accuracy of species recognition tasks. However, the black-box nature of these models presents a significant issue, as it is difficult to understand their internal decision-making processes, leading to concerns about their reliability and trustworthiness. This study addresses these issues by employing Explainable Artificial Intelligence (XAI) to enhance the transparency of deep learning models for bird species image classification. In this paper, a three-stage XAI-based approach is proposed, involving transfer learning, Local Interpretable Model-Agnostic Explanations (LIME), and Intersection over Union (IoU) scores to assess model performance. Six pretrained models are evaluated on the CUB 200-2011 dataset, with EfficientNetB0 achieving the highest accuracy (99.51%) and IoU score (0.43). Despite high accuracy, models such as InceptionResNetV2 and DenseNet201 showed lower IoU scores, raising trustworthiness concerns. This study underscores the importance of XAI in ensuring the transparency and reliability of Artificial Intelligence (AI) models in ecological applications.

1. Introduction

Birds are considered an indicator of the health of the ecological environment. They are indicators of biodiversity monitoring and reflect the strengths and weaknesses of human ecosystems. They play a crucial role in maintaining the balance of natural ecosystems (Kovařík et al., 2021; Wu et al., 2022). However, technological advancement and social progress have severely affected bird habitats (Lin et al., 2023; Aksoy et al., 2022), leading to a sharp decline in the diversity of bird species, with some species nearing extinction. Therefore, protecting birds has become a necessary and urgent task as it can help natural resource managers take appropriate conservation measures. Traditional methods of detecting, recognizing and evaluating the richness of bird species include mist nets, point counts, and transect counts (Fischer et al., 2023). However, these traditional methods are generally costly, biased and unable to collect data over long periods and large areas (Zhang et al., 2024; Wimmer et al., 2013).

Deep learning has revolutionized ecological research in recent years (Pichler and Hartig, 2023; Christin et al., 2019; Perry et al., 2022; Ryo, 2024). By analyzing vast amounts of bird image data, deep learning models can effectively identify different bird species based on their visual characteristics. This capability offers a significant advantage over traditional methods that rely on human expertise for bird identification, allowing for faster, more automated, and potentially more accurate ecological surveys (Kumar and Kondaveerti, 2023). However, despite achieving high accuracy, the internal working and decision making process of these models remains black-box, raising concerns about their reliability and trustworthiness (Minh et al., 2022). The internal mechanisms of deep learning models are difficult to understand due to their complex structure, often referred to as black boxes (Arrieta et al., 2020). This lack of transparency can result in undetected faults within the model and can perpetuate implicit biases in the training data, resulting in erroneous, unreliable, and undesired results (Antoniadi et al., 2021).

Explainable Artificial Intelligence (XAI) methods emerge as a vital response to the challenges posed by the blackbox nature of deep learning models (Ryo et al., 2021; Samek and Müller, 2019). XAI techniques aim to improve the model interpretability, enabling users to gain insight into the decision-making process of these models. Techniques

* Corresponding author.
E-mail addresses: samparthi.21phd7024@vitap.ac.in (S.V.S. Kumar), kishan.kondaveeti@gmail.com (H.K. Kondaveeti).

https://doi.org/10.1016/j.ecolind.2024.112886
Received 29 May 2024; Received in revised form 12 November 2024; Accepted 21 November 2024
Available online 7 December 2024

Abbreviations

Abbreviation	Description
AI	Artificial Intelligence
CNN	Convolutional Neural Network
GAN	Generative Adversarial Networks
Grad-CAM	Gradient-weighted Class Activation Mapping
IoU	Intersection over Union
LIME	Local Interpretable Model-Agnostic Explanations
M-CNN	Mask Convolutional Neural Network
SHAP	SHapley Additive ExPlanations
XAI	Explainable Artificial Intelligence

such as LIME (Aldughayfiq et al., 2023) generate local and global explanations for the predictions of deep learning models. These explanations highlight the crucial image regions or features that significantly influence the decision making process of models, providing valuable information on how decisions are made (Szandała, 2023). LIME also provides feature importance scores, highlighting the features that contribute the most to a particular prediction. Through such methods, users can better understand the rationale behind model outputs, facilitate informed decision making, and foster trust in Artificial Intelligence (AI) systems.

Using XAI in bird species image classification with deep learning models goes beyond research and conservation purposes. It has additional benefits, allowing enthusiasts and the general public to learn about automatic bird species image classification and how AI systems recognize them. Moreover, XAI helps identify potential biases in the training data or model predictions, ensuring fairness and providing opportunities to rectify them for enhanced bird species image classification.

The novel contributions of this work are as follows.

- A multilevel approach has been proposed to evaluate the reliability and trustworthiness of deep learning models for bird species image classification, in addition to classification accuracy.
- LIME method has been used to generate feature heatmap images to identify key features that influence model decisions.
- Widely used pre-trained models are considered to leverage their learned features for accurate bird image classification.
- A new metric IoU score has been introduced to quantify the degree of reliability of a model.

The remainder of this paper is structured as follows. Section 1 provides an introduction to the problem and discusses the contributions of this work. Section 2 reviews studies on bird species recognition, the necessity of XAI, and the reasons for choosing LIME over other XAI methods. In Section 3, we describe the methodology used in this study. Section 4 presents the results and their analysis. Section 5 covers related discussions and highlights the limitations of the study. Finally, Section 6 presents concluding remarks and suggests future research directions.

2. Scoping

The classification of bird species from images has gained significant attention in recent years, with transfer learning techniques emerging as a key focus area. The effectiveness of pretrained models, such as EfficientNetB5 and Inception-ResNet-v2, has been demonstrated through various studies (Huang and Basanta, 2021; Mochurad and Svystovych, 2024). Another category of works focused on the development of effective architectures which incorporate skip connections, which significantly improve feature extraction capacity in CNNs (Huang and Basanta, 2019; Farman et al., 2023).

In some studies, mobile applications have been developed that utilize CNN models and transfer learning to classify bird species with intuitive user interfaces (Van Horn et al., 2015; Choe et al., 2020; Gupta et al., 2021). Moreover, fine-grained classification methods have gained prominence, employing innovative techniques such as attention mechanisms and decoupled knowledge distillation to enhance model efficiency and improve accuracy (Wang et al., 2023; Ge et al., 2016; Wei et al., 2018). Furthermore, the integration of hyperparameter optimization techniques has proven beneficial in bird species recognition models, combining manual and random searches to fine-tune model parameters effectively (Kumar and Kondaveeti, 2024).

All of these studies focused on developing efficient and accurate deep learning models for bird species recognition, but they overlook whether the models are using the correct features for decision making. Relying on irrelevant features can lead to misclassifications, negatively impacting ecological conservation and biodiversity research. Furthermore, the lack of transparency in deep learning models complicates the interpretability of their predictions, making it difficult for users to understand the rationale behind classifications. This absence of clarity raises concerns about the reliability and trustworthiness of these models, especially in critical applications related to species identification and habitat preservation.

To ensure trust and reliability in deep learning models, it is crucial to evaluate their decision-making capabilities and underlying rationale (Araujo et al., 2020). Techniques such as SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017), Layerwise Relevance Propagation (LRP) (Bach et al., 2015), and Gradient-weighted Class Activation Mapping (Grad-CAM) (Selvaraju et al., 2017) are commonly used to provide visual explanations of model behavior by highlighting key features and their importance through heatmaps. Among these methods, SHAP is computationally intensive. LRP provides intuitive pixel-level insights but sometimes results in ambiguous interpretations in complex models. Grad-CAM suffers from localization issues, potentially highlighting large regions rather than specific features. In contrast, LIME is model-agnostic and excels at providing local explanations for individual predictions, making it user-friendly and adaptable. As LIME effectively estimates feature importance, we used it for identifying and masking important features based on feature importance.

In this study, we proposed a novel approach to implement XAI based model approach for reliability and trustworthiness of bird species image classification. Our methodology involves the use of image-LIME (Mathworks, 2024), a version of the LIME algorithm customized for image data, to visualize the key features selected by the models for decision making. Our goal is to improve the transparency and reliability of our model predictions while gaining insight into the features that contribute to accurate bird species image classification.

3. Methodology

The proposed methodology contains three stages. In stage 1, six widely used pretrained models and a benchmark dataset for bird species

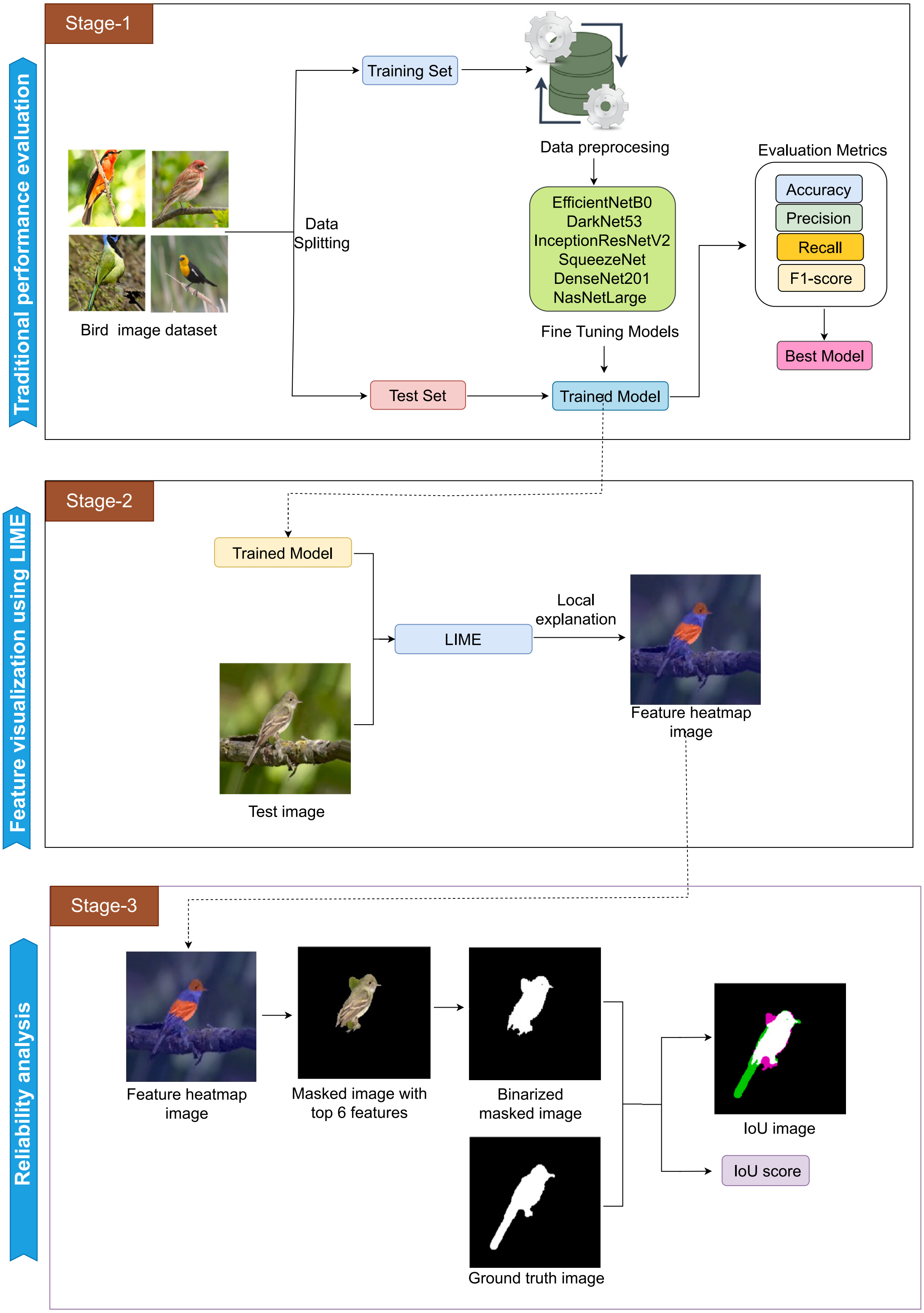

Fig. 1. Three stage approach proposed in the current study.

classification CUB 200-2011 is considered. The data set is divided into training and testing. DL models are trained on the training set and evaluated using standard metrics to assess their effectiveness. In stage 2, the LIME algorithm is employed to visualize the predictions of the DL models by identifying the important features of the image and regions that influenced their decisions. This interpretability analysis served two purposes: to evaluate model image recognition capabilities by assessing their reliance on meaningful visual cues and to analyze the completeness of the learned features for each bird species label using LIME visualizations. In stage 3, the degree of reliability of the DL models is identified using the quantitative metric IoU score.

The proposed methodology involves in training DL models on the CUB 200-2011 dataset (Wah et al., 2011) and measuring their performance using traditional metrics, interpreting model predictions using

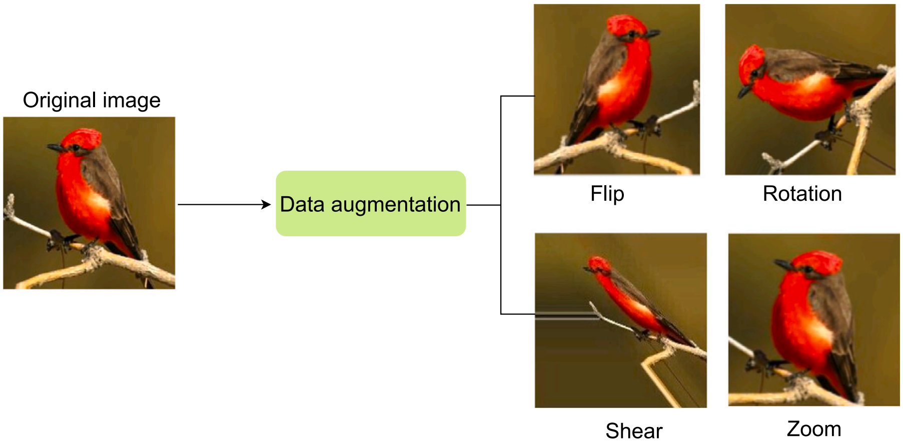

Fig. 2. Visualization depicting various data augmentation techniques employed in this study.

Table 1
Selected deep learning models information.

Pre-trained models	Parameters	Depth	Input image size
EfficientNetB0	5.31 million	82	224 × 224 × 3
DarkNet53	41.6 million	53	256 × 256 × 3
InceptionResNetV2	55.9 million	164	299 × 299 × 3
SqueezeNet	1.24 million	18	227 × 227 × 3
DenseNet201	20.0 million	201	224 × 224 × 3
NasNetLarge	88.9 million	744	331 × 331 × 3

LIME, and evaluating reliability using the IoU score. The methodology followed in this study is shown in Fig. 1. A detailed description of the methodology is provided in the following subsections.

3.1. Step 1 — Dataset description

The CUB 200-2011 dataset has become a fundamental resource for researchers for bird image classification. The data set consists of 200 bird species with 40–60 images per class, resulting in a grand total of 11,788 bird images. Each represents a bird in its natural environment, which presents challenges due to variations in pose, lighting, and background. Despite these variations, the dataset provides valuable annotations for each image.

3.2. Step 2 — Data prepocessing

The CUB 200-2011 dataset comprises 11,788 images and is used to train deep learning models. Due to the limited size of the dataset, data augmentation is a technique used to enhance the training performance of deep learning models by increasing the amount of available data. This involves creating synthetic data from existing data points through various transformations. In this study, four augmentation approaches are employed, such as shear, flip, zoom, and rotation. The data augmentation process resulted in the generation of 40,000 images, each class 200 images. For visual examples of these techniques are shown in Fig. 2.

The images are resized to match the standard input size requirements of the respective deep learning models used in the study. Detailed information on these models is presented in Table 1. This step ensures that the input images are compatible with the architectures of these pre-trained models, allowing for efficient processing and feature extraction.

3.3. Step 3 — Selection of pretrained models

Six widely used pretrained deep learning models, EfficientNetB0 (Tan and Le, 2019), DarkNet53 (Redmon and Farhadi, 2018), InceptionResNetV2 (Szegedy et al., 2017), SqueezeNet (Iandola et al., 2016), DenseNet201 (Huang et al., 2017), and NasNetLarge (Zoph et al., 2018), are considered in this study. These models encompass a diverse range of architectural designs and employ varying numbers and types of deep neural network layers. Although these models differ in their specific architectures, they share a common trait: increased depth and complexity generally lead to higher accuracy on large-scale datasets.

3.4. Step 4 — Training and fine-tuning the selected pretrained models

Transfer learning is a technique in deep learning in which a pretrained model, originally trained on a large dataset, is trained on another related task specific dataset and fine-tuned. This approach offers several advantages, including reduced training time, improved performance with limited data, and the ability to leverage the knowledge captured by the pretrained model.

EfficientNetB0, InceptionResNetV2, DenseNet201, DarkNet53, NASNetLarge, and SqueezeNet are the six pre-trained models used in the study. These models are initially trained on ImageNet (Krizhevsky et al., 2012), where the initial layers captured basic features such as shapes, colors, and patterns, while the final layers learned task-specific features. During transfer learning, for all models, new layers are added in place of these last three layers named a new learnable layer, a new softmax layer, and a new classification layer, where the initial layers are frozen and the new layers are fine-tuned using the CUB 200-2011 data set. The fewer newly added layers learned quickly during the fine-tuning process. A visual representation of the transfer learning process is illustrated in Fig. 3.

Optimal hyperparameter selection is crucial in transfer learning as it ensures that the model adapts effectively to the new task, maximizing its performance while avoiding overfitting. In this study, we carefully selected the hyperparameters, based on previous experimental results and domain knowledge (Kondaveeti et al., 2023a,b; Kumar and Kondaveeti, 2024), to optimize the performance of deep learning models. We used Adam optimizer with a learning rate of 0.0001 (Moreira and Fiesler, 1995) and a batch size of 32 (Luo et al., 2018), and categorical cross-entropy (Ho and Wookey, 2019) as the loss function. Categorical cross-entropy minimizes classification error by comparing predicted probabilities. The Adam optimizer is a widely used algorithm for training deep learning models (Soydaner, 2020; Huang and Basanta, 2021). Adam optimizer tunes the learning rate for each model parameter, improving training efficiency (Kingma, 2014; Jais et al., 2019).

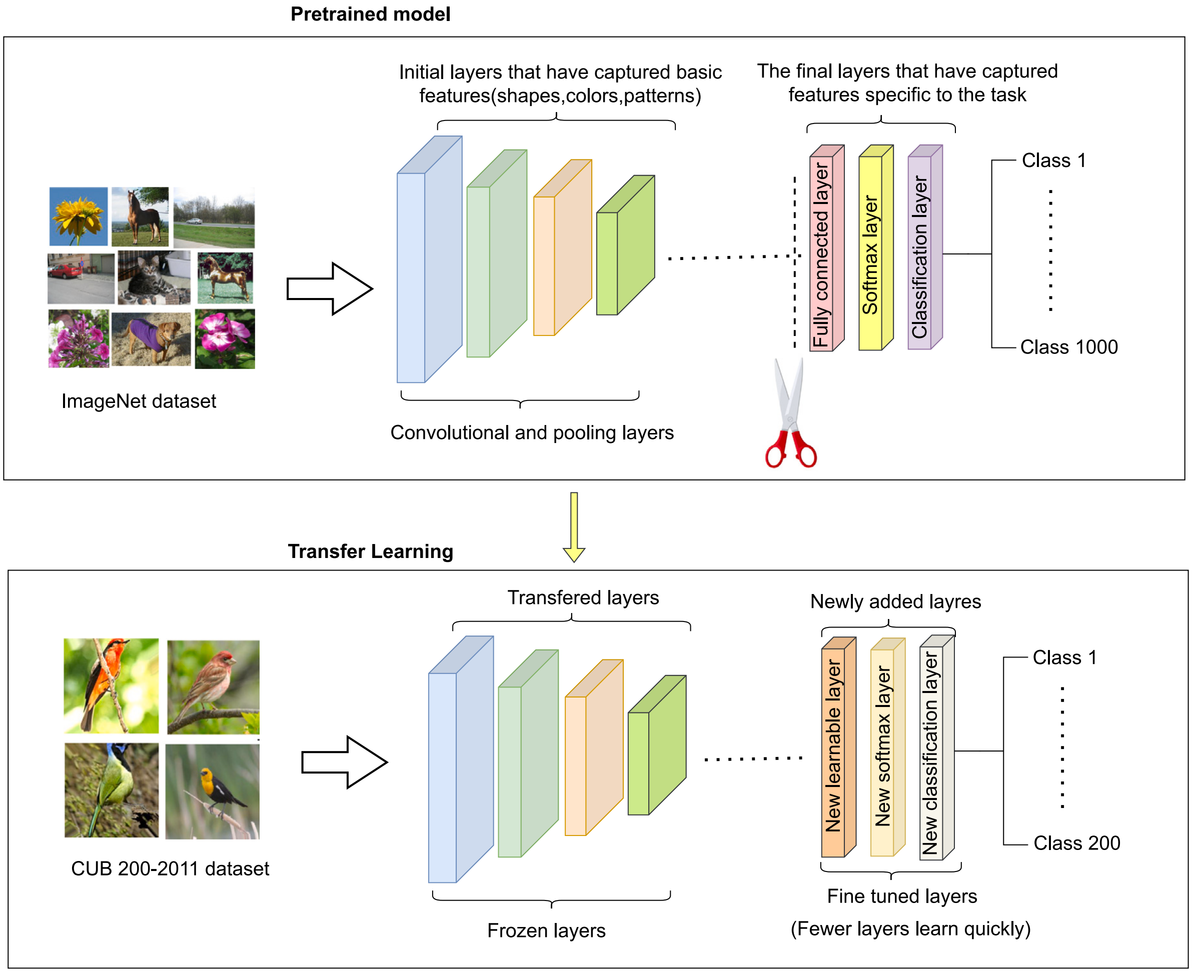

Fig. 3. Transfer learning approach.

The selected six pre-trained models are fine-tuned for 30 epochs on both the training and test sets. To avoid overfitting, the dropout regularization technique (Srivastava et al., 2014) is applied during the fine-tuning process. Dropout reduces overfitting by randomly deactivating neurons during training, thereby improving model generalization. While retaining the original layers of the pretrained models, modifications are made solely to the output layers. This approach not only helped prevent overfitting but also facilitated faster convergence. Table 1 provides detailed information about the fixed parameters and the basic structure of the selected models.

3.5. Step 5 — Evaluation of deep learning models based on classification performance metrics

In Stage-1, our objective is to identify an efficient method to classify images of birds. After data augmentation, this data set contains 40,000 images and serves as a basis for training six pretrained models using transfer learning. The models chosen has previously shown exceptional performance on the ImageNet dataset. The study involved freezing the convolutional layers of these pre-trained models and adding new fully connected layers to adapt them to the specific task of bird species image classification. The experiments are carried out by dividing the data set into training and testing sets in three different ratios 50:50, 60:40, and 70:30.

Four key metrics are used for the evaluation of classification performance. Accuracy, which measures overall prediction correctness; precision, which assesses the ability to correctly identify true positive bird instances; recall assesses the ability to find all true positive bird instances; and the F1 score, which is the harmonic mean of precision and recall, balancing their trade-off. Table 2 shows the details of the performance metrics.

3.6. Step 6 — LIME approach

LIME is used to explain the predictions of any classifier in a clear way. LIME creates an interpretable model that closely approximates the behavior of the complex machine learning model locally around a prediction (Ribeiro et al., 2016). LIME identifies the features that have the greatest impact on a prediction by perturbing the neighborhood of a sample and observing any changes in the prediction (Zafar and Khan, 2021). LIME focuses on generating explanations that make sense for individual predictions, even if only a small number of variables matter locally compared to the overall global context.

3.7. Step 7 — Feature heatmap image generation using LIME

In stage 2, LIME (Ribeiro et al., 2016) is used to visualize and understand the decision-making process of the model at a local level by highlighting the important regions or features in an image that contribute to the prediction of the models. The LIME interpretations of the top “n” features considered selected models for an image from our test data set. We then used LIME-generated feature heatmaps to mask the images, revealing only the most crucial 6, 8, and 10 features. This process helps us to understand the model decision-making process clearly. Fig. 4 illustrates this with six selected models: (a) the original image of the Kentucky Warbler bird, (b) the feature heatmap images

Table 2
Model performance metrics.

Metrics	Interpretation	Mathematical expression	Range	Remarks
Accuracy	Ratio of correctly predicted instances to the total number of instances	Accuracy $= \frac{TP+TN}{TP+FP+TN+FN} \times 100$	$[0, 100]$	Values closer to 100 indicate better performance.
Precision	Ratio of true positives to the sum of true positives and false positives	Precision $= \frac{TP}{TP+FP} \times 100$	$[0, 100]$	Values closer to 100 indicate a better ability to classify positive instances correctly.
Recall	Ratio of true positives to the sum of true positives and false negatives	Recall $= \frac{TP}{TP+FN} \times 100$	$[0, 100]$	Values closer to 100 indicate a better ability to capture positive instances.
F1-score	Harmonic mean of precision and recall	F1-score $= \frac{2\cdot\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}} \times 100$	$[0, 100]$	Values closer to 100 indicate a better balance between precision and recall.

generated by LIME, (c) the masked image with the top 6 features, (d) the image masked with the top eight features, and (e) the masked image with the top 10 features.

When examining the top 6 and 8 most important features identified by the selected models using LIME, crucial parts of the birds, such as the beak, eyes, and wings, are not adequately highlighted. However, when considering the top 10 features, these vital regions are effectively emphasized, allowing for better recognition of the birds key anatomical features.

Among the models evaluated, EfficientNetB0, InceptionResNetV2, DenseNet201, and NasNetLarge demonstrate the best identification of bird parts when focusing on the 10 most relevant features. In this analysis, we refrain from using quantitative metrics, as the assessment is primarily based on the visual inspection of the LIME explanations rather than numerical analysis, aiming to qualitatively evaluate the models ability to focus on the appropriate regions for accurate bird species classification.

3.8. Step 8 — Model reliability analysis

The final stage of the study involves a comprehensive comparison of the models that have demonstrated high classification accuracy in the first phase and effective feature extraction using LIME in the second phase. The similarity metric, IoU, is used to evaluate how well these models focus on relevant features in bird species images. LIME visualizations are used to understand the features that each model utilizes to make predictions. The efficiency of feature extraction is measured by the similarity between the model identified features and the actual ground truth features present in the images. A metric Intersection over Union (IoU) score is used to quantitatively compare binary masked image and ground truth image and quantify the degree of overlap between the features selected, IoU enables the assessment of the models ability to extract relevant features for accurate bird species classification. Figs. 5 and 6 demonstrate comparative quantitative analysis of the feature selection process and the ability of the models to identify the most relevant features for the accurate classification of bird species. The comparison between the masked binarized image and the ground truth, along with the IoU metric, helps assess the reliability of the model feature selection. IoU measures the similarity between two sets by calculating the ratio of the intersection masked binary image and ground truth area to their union, providing a quantitative assessment of feature overlap. The IoU ranges from 0 to 1, where 0 indicates no similarity and 1 indicates perfect similarity. Eq. (1) compares the overlap of ground truth images (GT(i,j)) and masked binary images (MB(i,j)).

$$\text{IoU}(GT, MB) = \frac{\sum_{j=1}^{N}\sum_{i=1}^{M}(GT(i,j) \cap MB(i,j))}{\sum_{j=1}^{N}\sum_{i=1}^{M}(GT(i,j) \cup MB(i,j))} \tag{1}$$

4. Experimental setup and result analysis

The experiments are carried out on a system equipped with a NVIDIA GeForce GTX 1080 Ti graphics card, 64 GB of RAM, and an Intel (R) Xeon (R) W-2125 219 processor operating at a clock speed of 4.00 GHz. The experiments are performed using the MATLAB R2024a platform, leveraging its advanced computational capabilities and libraries. This high-performance computing environment ensured efficient training and evaluation of the deep learning models, allowing thorough experimentation and analysis of the proposed approaches for bird species classification.

The results displayed in Table 3 represent the performance comparison of selected DL models with conventional evaluation metrics on different data partitions.

Among the selected models, EfficientNetB0 achieved the highest accuracy of 99.51%, along with impressive precision 99. 39%, recall 99. 38% and the F1 score 99. 38%, indicating its exceptional performance in accurately classifying bird species. SqueezeNet and InceptionResNetV2 also demonstrated remarkable results, with accuracy scores above 99% and consistent performance in other metrics. Although DenseNet201 and DarkNet53 exhibited relatively lower scores compared to the top performing models, their accuracy values are 96.53% and 96.09%, respectively. These results indicate that EfficientNetB0 shows strong performance in bird species image classification.

Table 4 presents the IoU values for various CNN models based on the number of features considered. EfficientNetB0, InceptionResNetV2, DenseNet201, DarkNet53, NasNetLarge, and SqueezeNet are evaluated with 6, 8, and 10 features, showing increasing IoU values as the number of features increases. The average IoU across features is also considered, which offers insight into how well each model focuses on relevant features for accurate predictions. This comprehensive evaluation allows researchers to identify models that not only achieve high classification accuracy but also demonstrate reliable feature selection. EfficientNetB0 exhibits a high average IoU of 0.43, indicating that it consistently identifies relevant features in different feature sets. In contrast, DenseNet201 shows the lowest average IoU of 0.34, demonstrating less reliable feature selection. The EfficientNetB0 model achieved an impressive accuracy of 99.51% and a high IoU score of 0.43, demonstrating its efficiency and reliability. Although InceptionResNetV2 and DenseNet201 models achieved a high accuracy of 98.96% and 96.53%, respectively, their lower IoU scores of 0.36 and 0.34 raise concerns about their trustworthiness. SqueezeNet achieved 99.21% accuracy, which is higher than that of NasNetLarge 98.64%. However, the IoU score of 0.42 for NasNetLarge indicates greater reliability compared to SqueezeNet with 0.40.

5. Discussion

The study proposes a three-stage approach using XAI through LIME, demonstrating promising results in evaluating selected deep learning models in terms of efficiency and reliability. In the first stage, the

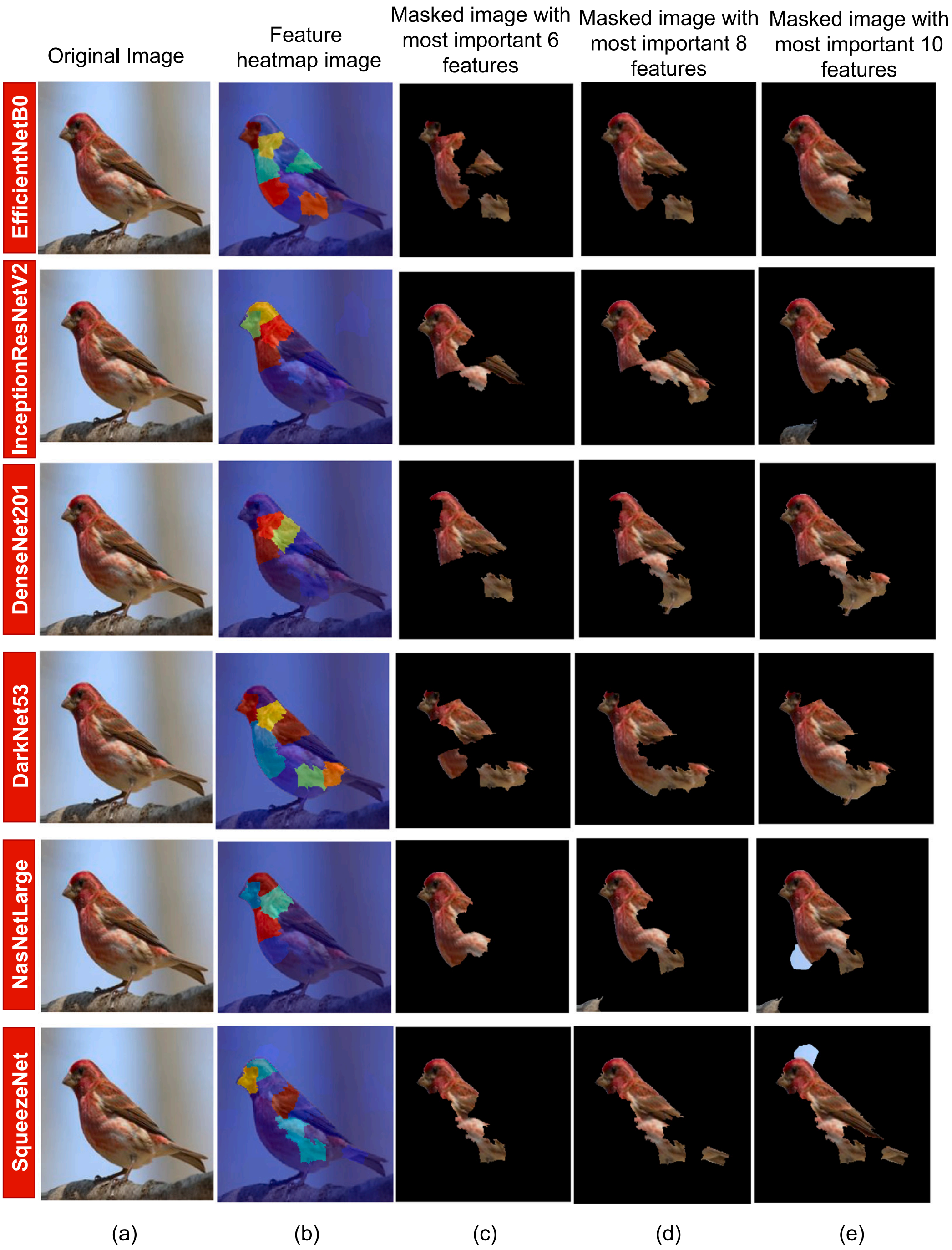
Original Image
Feature heatmap image
Masked image with most important 6 features
Masked image with most important 8 features
Masked image with most important 10 features
EfficientNetB0
InceptionResNetV2
DenseNet201
DarkNet53
NasNetLarge
SqueezeNet
(a)
(b)
(c)
(d)
(e)

Fig. 4. (a) Original image of the bird Kentucky Warbler; (b) Feature heatmaps images generated using LIME; (c) Masked image with most important 6 features; (d) Masked image with most important 8 features; (e) Masked image with most important 10 features.

results indicate that the EfficientNetB0 model achieves the highest accuracy. However, in the second phase, utilizing XAI with LIME to provide clear explanations for the inner workings of the black box models is crucial for in-depth assessments and ensuring reliability in practical applications. The third stage reveals that there is more IoU despite the higher recognition results of EfficientNetB0. This finding confirms that, while the accuracy assessment yields the best results, it does not necessarily indicate the reliable and trustworthy model. Although the SqueezeNet model achieves testing accuracy greater than 99%, the learned features are not significant. Consequently, the model may misclassify images in real-world scenarios if the image is incomplete or the background differs. Through LIME, it becomes possible to clearly compare the performance of the models and ensure that the model learns essential features from the XAI-based approach, which can improve the model and refine the data set in a particular direction to enhance the reliability of the models in the research.

Although our research demonstrates promising results in implementing an XAI model (LIME) for bird species classification, there are several limitations that need to be addressed to enhance the robustness and applicability of the findings. The study considers a limited number of classes for bird species, which may not capture the full diversity of bird species. Expanding the number of classes could provide a more comprehensive evaluation of the models. The number of pretrained models used in this study is limited. Including a wider variety of models could offer a more thorough comparison and identify the most effective architectures for this task. Another significant limitation is

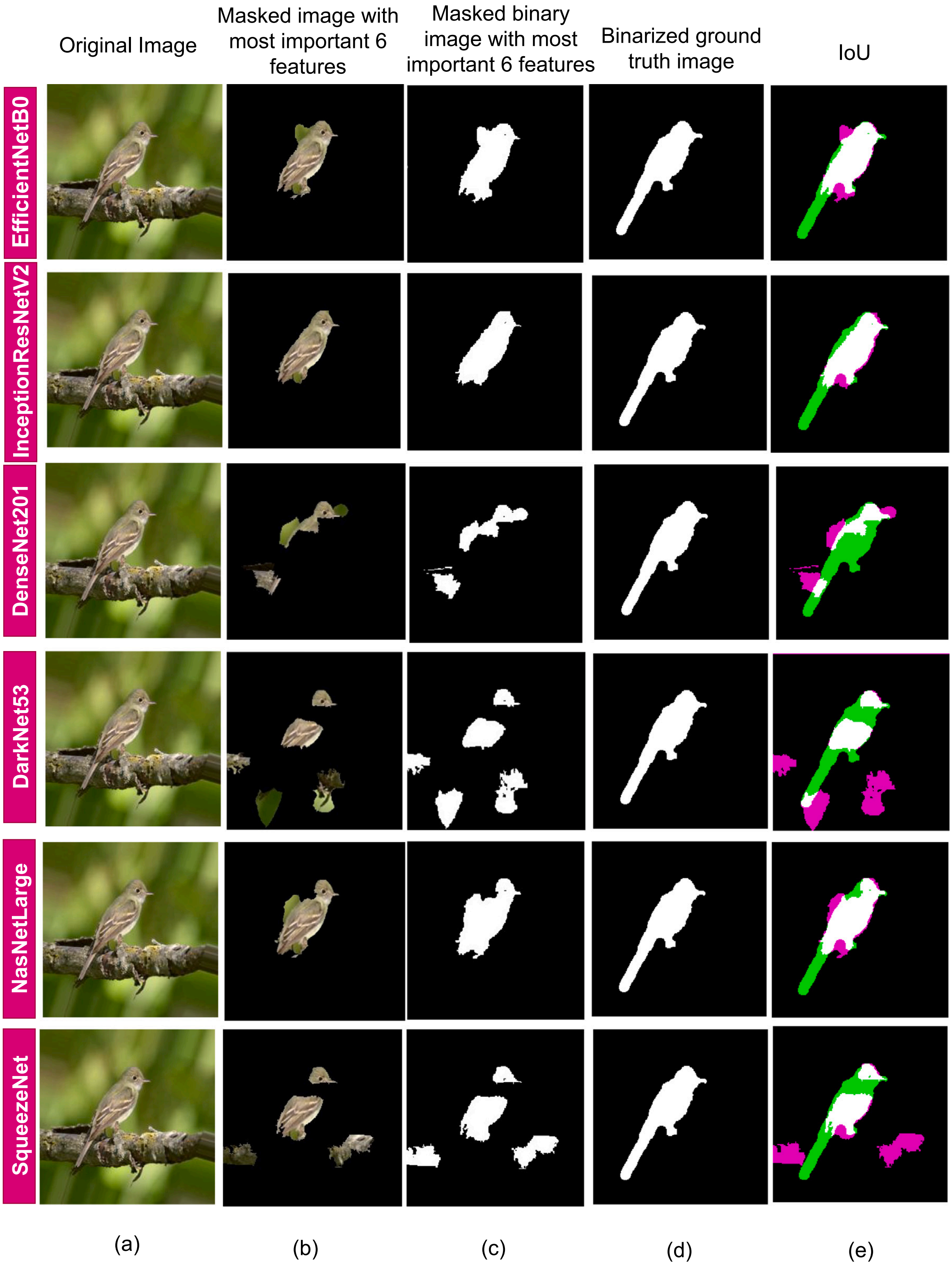
Original Image
Masked image with most important 6 features
Masked binary image with most important 6 features
Binarized ground truth image
IoU
EfficientNetB0
InceptionResNetV2
DenseNet201
DarkNet53
NasNetLarge
SqueezeNet
(a)
(b)
(c)
(d)
(e)

Fig. 5. IoU for top 6 features: (a) Original image of the bird Purple Finch; (b) Masked image with most important top 6 features; (c) Masked binary image with top 6 features; (d) Binarized ground truth image; (e) IoU image; Color coding: Pink color pixels represents the masked binary image, Green color pixels represents the ground truth image, White color pixels represents the overlapping area of both.

the lack of information on the optimal number of features that should be considered for analysis. Our experiment does not determine the optimal number of features to focus on, which could affect the interpretability and performance of the models. Furthermore, we did not consider hyperparameter optimization, which could potentially improve the accuracy and efficiency of the models. Our study uses LIME as the XAI technique. Exploring other XAI methods such as Gradient-weighted Class Activation Mapping (Grad-CAM) and SHapley Additive ExPlanations (SHAP) could provide different insights and potentially more accurate explanations of the model decision-making processes.

We only used a publicly available CUB 200-2011 dataset. Incorporating more diverse and extensive datasets, including proprietary or less

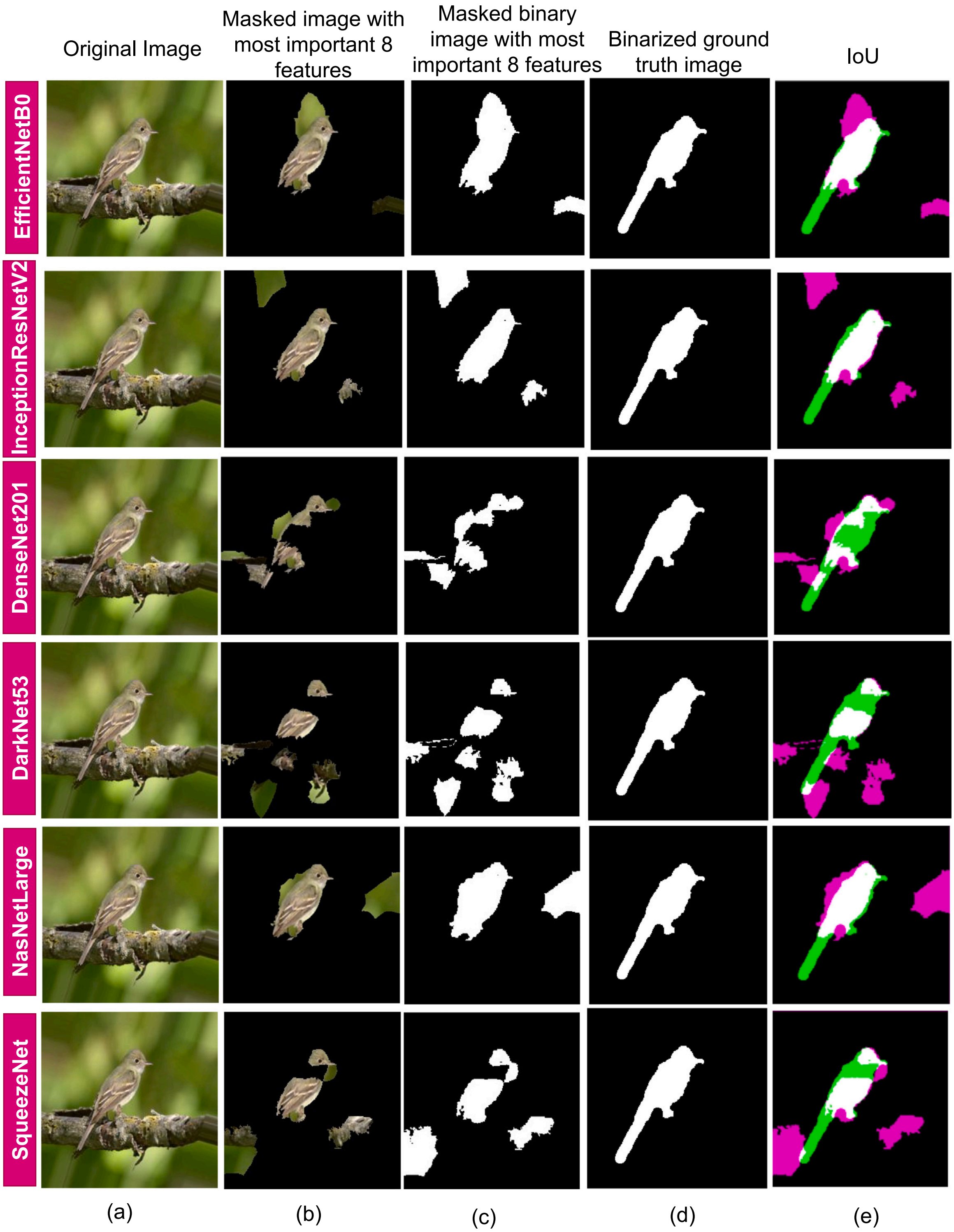
Original Image
Masked image with most important 8 features
Masked binary image with most important 8 features
Binarized ground truth image
IoU
EfficientNetB0
InceptionResNetV2
DenseNet201
DarkNet53
NasNetLarge
SqueezeNet
(a)
(b)
(c)
(d)
(e)

Fig. 6. IoU for top 8 features: (a) Original image of the bird Purple Finch; (b) Masked image with most important top 8 features; (c) Masked binary image with top 8 features; (d) Binarized ground truth image; (e) IoU image; Color coding: Pink color pixels represents the masked binary image, Green color pixels represents the ground truth image, White color pixels represents the overlapping area of both.

commonly used ones, could improve the generalizability of the results. The research primarily uses the IoU as a quantitative metric, which may not capture all aspects of feature importance to identify model reliability. The generation of ground truth images, a crucial step in the evaluation process, can be a challenging task and may introduce uncertainties in the analysis.

Our experiments are not conducted on real-time data, which could affect the practical applicability of the models in real-world scenarios. Conducting real-time experiments would provide a better understanding of how models perform in dynamic and unpredictable environments.

Data augmentation techniques such as zoom, shear, rotation, and flip are applied to enhance the dataset. Although these techniques are useful, exploring other augmentation methods could further improve the robustness of the models and generalization capabilities. It is important to note that these limitations do not diminish the significance

Table 3
Performance comparison of selected DL models with conventional evaluation metrics on different data partitions.

Data partition	Pre-trained models	Accuracy	Precision	Recall	F1 score
50:50	EfficientNetB0	99.42	99.34	99.33	99.33
	InceptionResNetV2	98.83	98.84	98.83	98.83
	DenseNet201	96.33	96.43	96.33	96.31
	DarkNet53	94.67	95.19	94.66	94.71
	NasNetLarge	99.33	99.34	99.33	99.33
	SqueezeNet	99.26	99.25	99.22	99.29
60:40	EfficientNetB0	99.57	99.59	99.58	99.41
	InceptionResNetV2	98.33	98.39	98.33	98.33
	DenseNet201	96.04	96.09	96.04	96.02
	DarkNet53	97.50	97.60	97.50	97.50
	NasNetLarge	97.71	97.77	97.70	97.69
	SqueezeNet	99.21	99.19	99.23	99.31
70:30	EfficientNetB0	99.53	99.25	99.24	99.41
	InceptionResNetV2	99.72	99.72	99.72	99.72
	DenseNet201	97.22	97.32	97.22	97.21
	DarkNet53	96.11	96.51	96.11	96.16
	NasNetLarge	98.89	98.93	98.86	98.87
	SqueezeNet	99.17	99.18	99.16	98.99
Average	EfficientNetB0	99.51	99.39	99.38	99.38
	InceptionResNetV2	98.96	98.98	98.96	98.96
	DenseNet201	96.53	96.61	96.53	96.51
	DarkNet53	96.09	96.43	96.09	96.12
	NasNetLarge	98.64	98.68	98.63	98.63
	SqueezeNet	99.21	99.21	99.20	99.20

Table 4
IoU score of the selected pretrained models.

CNN models	Intersection over union (IoU) score			
	6 features	8 features	10 features	Average IoU score
EfficientNetB0	0.38	0.44	0.45	0.43
InceptionResNetV2	0.34	0.36	0.37	0.36
DenseNet201	0.31	0.34	0.35	0.34
DarkNet53	0.35	0.38	0.40	0.38
NasNetLarge	0.38	0.42	0.44	0.42
SqueezeNet	0.36	0.41	0.43	0.40

of this research but highlight potential areas for further exploration and improvement in future studies.

6. Conclusion and future work

This study uses explainable and interpretable techniques to recognize bird species from images to understand how convolutional layers within deep learning models contribute to the classification process. We used the LIME methodology to achieve interpretability and trained six transfer learning models on the CUB 200-2011 dataset. EfficientNetB0 demonstrated the highest accuracy of 99.51%, along with impressive precision of 99.39%, recall of 99.38%, and F1-score of 99.38%, indicating its exceptional performance in accurately classifying bird species. SqueezeNet and InceptionResNetV2 also showed remarkable results, with accuracy scores above 99% and consistent performance in other metrics. Although DenseNet201 and DarkNet53 exhibited relatively lower scores, their accuracy levels of 96.53% and 96.09%, respectively.

In stage 2, the LIME algorithm is used to explain the predictions of the models by identifying the important image features and regions that influenced their decisions. This interpretability analysis served two purposes: evaluating the image recognition capabilities of models by assessing their reliance on meaningful visual cues and analyzing the learned features for each bird species label using LIME visualizations.

In stage 3, the best model is identified using the quantitative metric IoU, which measures the similarity between two sets. These results highlight EfficientNetB0 strong performance in image classification, showcasing its effectiveness in image recognition tasks. Overall, EfficientNetB0 stands out as the most trustworthy model, combining high classification accuracy with effective feature extraction. By incorporating XAI through LIME, the proposed methodology becomes more reliable and interpretable, providing valuable insights into the feature extraction process.

Although this research has a significant impact on improving bird species image classification using explainable AI techniques, there are several avenues for future work to further enhance the reliability, efficiency, and applicability. The research can be extended by incorporating other XAI techniques, such as Grad-CAM and SHAP, to compare their effectiveness with LIME and determine which method provides the most reliable and interpretable results.

Implementing our work on real-time datasets will test the practical applicability of our models and ensure that they perform well in dynamic environments. Conducting experiments on additional similarity metrics will provide a deeper evaluation of the reliability of the model.lity. To save time and reduce manual effort, future research may explore automatic image segmentation methods, which can significantly streamline the preprocessing phase. Experiments with advanced deep learning models and newer versions of selected models should be carried out to assess their potential to improve classification accuracy and feature extraction. Future studies should also consider computation time and memory occupation to ensure that the models are efficient and scalable. Exploring advanced augmentation techniques, such as Generative Adversarial Networks (GAN), can enhance the diversity and robustness of the training dataset, further mitigating overfitting. Furthermore, experimenting with real-time bird species datasets will provide valuable insight into the performance of the model in practical applications.

CRediT authorship contribution statement

Samparthi V.S. Kumar: Writing – review & editing, Writing – original draft. **Hari Kishan Kondaveeti:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Conceptualization.

Code availability

The code developed and used in this study has been uploaded and is available on GitHub and MATLAB. These repositories include all the necessary scripts and instructions required to replicate the results and reuse them in future studies.

The repository can be accessed via the following links:
Github Link: https://shorturl.at/Github_Bird_XAI
MATLAB File Exchange Link: https://shorturl.at/Matlab_Bird_XAI

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The dataset used in this study is publicly available. The repository can be accessed via the following link:https://www.kaggle.com/datasets/veeralakrishna/200-bird-species-with-11788-images .

References

Aksoy, T., Dabanli, A., Cetin, M., Senyel Kurkcuoglu, M.A., Cengiz, A.E., Cabuk, S.N., Agacsapan, B., Cabuk, A., 2022. Evaluation of comparing urban area land use change with urban atlas and CORINE data. Environ. Sci. Pollut. Res. 29 (19), 28995–29015.

Aldughayfiq, B., Ashfaq, F., Jhanjhi, N., Humayun, M., 2023. Explainable AI for retinoblastoma diagnosis: interpreting deep learning models with LIME and SHAP. Diagnostics 13 (11), 1932.

Antoniadi, A.M., Du, Y., Guendouz, Y., Wei, L., Mazo, C., Becker, B.A., Mooney, C., 2021. Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: a systematic review. Appl. Sci. 11 (11), 5088.

Araujo, T., Helberger, N., Kruikemeier, S., De Vreese, C.H., 2020. In AI we trust? Perceptions about automated decision-making by artificial intelligence. AI Soc. 35 (3), 611–623.

Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al., 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion 58, 82–115.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., Samek, W., 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS One 10 (7), e0130140.

Choe, D., Choi, E., Kim, D.K., 2020. The real-time mobile application for classifying of endangered parrot species using the CNN models based on transfer learning. Mob. Inf. Syst. 2020 (1), 1475164.

Christin, S., Hervet, É., Lecomte, N., 2019. Applications for deep learning in ecology. Methods Ecol. Evol. 10 (10), 1632–1644.

Farman, H., Ahmed, S., Imran, M., Noureen, Z., Ahmed, M., 2023. Deep learning based bird species identification and classification using images. J. Comput. Biomed. Inf. 6 (01), 79–96.

Fischer, S., Edwards, A.C., Garnett, S.T., Whiteside, T.G., Weber, P., 2023. Drones and sound recorders increase the number of bird species identified: A combined surveys approach. Ecol. Inform. 74, 101988.

Ge, Z., Bewley, A., McCool, C., Corke, P., Upcroft, B., Sanderson, C., 2016. Fine-grained classification via mixture of deep convolutional neural networks. In: 2016 IEEE Winter Conference on Applications of Computer Vision. WACV, IEEE, pp. 1–6.

Gupta, D., et al., 2021. Mobile application for bird species identification using transfer learning. In: 2021 IEEE International Conference on Artificial Intelligence in Engineering and Technology. IICAIET, IEEE, pp. 1–6.

Ho, Y., Wookey, S., 2019. The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling. IEEE Access 8, 4806–4813.

Huang, Y.-P., Basanta, H., 2019. Bird image retrieval and recognition using a deep learning platform. IEEE Access 7, 66980–66989.

Huang, Y.-P., Basanta, H., 2021. Recognition of endemic bird species using deep learning models. Ieee Access 9, 102975–102984.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4700–4708.

Iandola, F.N., Han, S., Moskewicz, M.W., Ashraf, K., Dally, W.J., Keutzer, K., 2016. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and < 0.5 MB model size. arXiv preprint arXiv:1602.07360.

Jais, I.K.M., Ismail, A.R., Nisa, S.Q., 2019. Adam optimization algorithm for wide and deep neural network. Knowl. Eng. Data Sci. 2 (1), 41–46.

Kingma, D.P., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kondaveeti, H.K., Guturu, S.S.V., Praveen, K.J., Kumar, S.V., 2023a. A transfer learning approach to bird species recognition using MobileNetV2. In: 2023 7th International Conference on Intelligent Computing and Control Systems. ICICCS, IEEE, pp. 787–794.

Kondaveeti, H.K., Nithiyasri, P., Sri, B.S.L., Jessica, K.H., Kumar, S.V., Gopi, S.C., 2023b. Bird species recognition using deep learning. In: 2023 3rd International Conference on Artificial Intelligence and Signal Processing. AISP, IEEE, pp. 1–6.

Kovařík, P., Pechanec, V., Machar, I., Harmáček, J., Grim, T., 2021. Are birds reliable indicators of most valuable natural areas? Evaluation of special protection areas in the context of habitat protection. Ecol. Indic. 132, 108298.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. 25.

Kumar, S.V., Kondaveerti, H.K., 2023. A comparative study on deep learning techniques for bird species recognition. In: 2023 3rd International Conference on Intelligent Communication and Computational Techniques. ICCT, IEEE, pp. 1–6.

Kumar, S.V., Kondaveeti, H.K., 2024. Bird species recognition using transfer learning with a hybrid hyperparameter optimization scheme (HHOS). Ecol. Inform. 102510.

Lin, D.-L., Ko, J.C.-J., Amano, T., Hsu, C.-T., Fuller, R.A., Maron, M., Fan, M.-W., Pursner, S., Wu, T.-Y., Wu, S.-H., et al., 2023. Taiwan's Breeding Bird Survey reveals very few declining species. Ecol. Indic. 146, 109839.

Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. 30.

Luo, P., Wang, X., Shao, W., Peng, Z., 2018. Towards understanding regularization in batch normalization. arXiv preprint arXiv:1809.00846.

Mathworks, 2024. Understand Network Predictions Using LIME - MATLAB & Simulink - MathWorks India — in.mathworks.com. https://in.mathworks.com/help/deeplearning/ug/understand-network-predictions-using-lime.html. (Accessed 28-05-2024).

Minh, D., Wang, H.X., Li, Y.F., Nguyen, T.N., 2022. Explainable artificial intelligence: a comprehensive review. Artif. Intell. Rev. 1–66.

Mochurad, L., Svystovych, S., 2024. A new efficient classifier for bird classification based on transfer learning. J. Eng. 2024 (1), 8254130.

Moreira, M., Fiesler, E., 1995. Neural networks with adaptive learning rate and momentum terms.

Perry, G.L., Seidl, R., Bellvé, A.M., Rammer, W., 2022. An outlook for deep learning in ecosystem science. Ecosystems 25 (8), 1700–1718.

Pichler, M., Hartig, F., 2023. Machine learning and deep learning—A review for ecologists. Methods Ecol. Evol. 14 (4), 994–1016.

Redmon, J., Farhadi, A., 2018. Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767.

Ribeiro, M.T., Singh, S., Guestrin, C., 2016. " why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1135–1144.

Ryo, M., 2024. Ecology with artificial intelligence and machine learning in Asia: A historical perspective and emerging trends. Ecol. Res. 39 (1), 5–14.

Ryo, M., Angelov, B., Mammola, S., Kass, J.M., Benito, B.M., Hartig, F., 2021. Explainable artificial intelligence enhances the ecological interpretability of black-box species distribution models. Ecography 44 (2), 199–205.

Samek, W., Müller, K.-R., 2019. Towards explainable artificial intelligence. In: Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K., Müller, K.-R. (Eds.), Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Springer International Publishing, Cham, pp. 5–22. http://dx.doi.org/10.1007/978-3-030-28954-6_1.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 618–626.

Soydaner, D., 2020. A comparison of optimization algorithms for deep learning. Int. J. Pattern Recognit. Artif. Intell. 34 (13), 2052013.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15 (1), 1929–1958.

Szandała, T., 2023. Unlocking the black box of CNNs: Visualising the decision-making process with PRISM. Inform. Sci. 642, 119162.

Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A., 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 31, (1).

Tan, M., Le, Q., 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning. PMLR, pp. 6105–6114.

Van Horn, G., Branson, S., Farrell, R., Haber, S., Barry, J., Ipeirotis, P., Perona, P., Belongie, S., 2015. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 595–604.

Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S., 2011. The caltech-ucsd birds-200-2011 dataset. (Accessed 11-April-2023).

Wang, K., Yang, F., Chen, Z., Chen, Y., Zhang, Y., 2023. A fine-grained bird classification method based on attention and decoupled knowledge distillation. Animals 13 (2), 264.

Wei, X.-S., Xie, C.-W., Wu, J., Shen, C., 2018. Mask-CNN: Localizing parts and selecting descriptors for fine-grained bird species categorization. Pattern Recognit. 76, 704–714.

Wimmer, J., Towsey, M., Roe, P., Williamson, I., 2013. Sampling environmental acoustic recordings to determine bird species richness. Ecol. Appl. 23 (6), 1419–1428.

Wu, J., Jin, X., Wang, H., Feng, Z., 2022. Evaluating the supply-demand balance of cultural ecosystem services with budget expectation in Shenzhen, China. Ecol. Indic. 142, 109165.

Zafar, M.R., Khan, N., 2021. Deterministic local interpretable model-agnostic explanations for stable explainability. Mach. Learn. Knowl. Extr. 3 (3), 525–541.

Zhang, C., Jin, N., Xie, J., Hao, Z., 2024. CicadaNet: Deep learning based automatic cicada chorus filtering for improved long-term bird monitoring. Ecol. Indic. 158, 111423.

Zoph, B., Vasudevan, V., Shlens, J., Le, Q.V., 2018. Learning transferable architectures for scalable image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8697–8710.