Proceedings of 15th IOE Graduate Conference
Peer Reviewed
Year: 2024 Month: May Volume: 15
ISSN: 2350-8914 (Online), 2350-8906 (Print)

An Integrated Approach for Wildlife Recognition in Nepal Using Video Along With Audio

Zishan Siddique [a], Sandip Pariyar [b], Venus Bastola [c], Sushant Phagu [d], Binay Lal Shrestha [e], Pravin Sangroula [f]

[a, b, c, d, e, f] *Purwanchal Campus, IOE, Tribhuvan University, Nepal*
[a] 076bct095@ioepc.edu.np, [b] 076bct074@ioepc.edu.np, [c] 076bct091@ioepc.edu.np, [d] 076bct090@ioepc.edu.np, [d] binay@ioepc.edu.np, [d] pravin@ioepc.edu.np

Abstract

Wildlife around the world is in decline primarily due to the loss of habitat as well as the intersection of territory between humans and wild animals. Manual recognition of animals can be more accurate but will require exponentially greater resources both in terms of capital and labor making it unfeasible in large-scale deployment, especially for a country like Nepal. We have developed a system for the recognition and classification of wild animals using deep convolutional neural network (DCNN) architecture to aid conservation as well as a study of our ecological system. We used iNaturalist as our source for image data and Xeno-canto.org as the source for audio data. We were able to achieve an F1-Score of 86.12% on image data of 44 species of animals and an F1-Score of 88.1% on audio data of 23 species of birds all found in Nepal.

Keywords

Convolutional Neural Network, EfficientNet, Audio, Video

1. Introduction

Nepal is a land of unparalleled biodiversity with 208 species of mammals [1], 142 species of reptiles [2], and more than 873 species of birds [3]. The diverse wildlife, including endangered species found here, are testament to the country's commitment to conservation. However, with the increasing challenges posed by habitat loss, poaching, and climate change, preserving Nepal's wildlife heritage demands innovative and technologically advanced approaches. This project provides a basic template for creating solutions to bolster wildlife conservation efforts in Nepal.

Vision can be sufficient for some species that are of adequate size but vision alone can miss many opportunities for identification and classification. For this reason, we have used two models i.e. vision and audio for a broader scope. The paper is structured as follows:

- Earlier approaches for our objectives are discussed in Section 2.
- The datasets used and their meta information are explained in Section 3.
- Our approach for the proposed system is elaborated in Section 4.
- The results of our solution are explained in Section 5.
- Finally, the complete project is concluded with a brief paragraph in Section 6.

2. Background and Related Works

There have been many instances of work done for wildlife. In 2017, Nguyen et al.[4] worked on a monitoring system for animals in South-central Victoria, Australia with appreciable results. They used deep convolutional neural networks (DCNNs) to train a model on 3.2 million images of animals spanning 48 species from Tanzania's Serengeti National Park.

Gautam et al. 2023 [5] worked on the classification of birds based on audio data of 44 species found in Nepal. Their work was based on Mel Spectogram features which are often termed "pictures of a sound". Mel Spectogram and MFCC (Mel-Frequency Cepstral Coefficients) were used for feature extraction of audio data which was used to train a Convolutional Neural Network.

In 2021, Qi et al. [6] worked on image recognition in Animal Husbandry. They applied super pixel-based image segmentation and SIFT algorithm to complete image segmentation and feature extraction which was later modeled using Convolutional Neural Network and SVM for classification.

We propose an integrated system of using both audio and video. The audio is used as it is with some pre-processing and the video is broken into frames which are fed to the image classifier model.

3. Dataset Analysis

As two mediums, vision and audio are being used for classification of animals, two different datasets comprising of labeled images and labeled audio are required. ImageNet was our the initial choice for image data but due to the misrepresentation of species, the idea was dropped, and iNaturalist was chosen as our source for image data for its reliability and proper representation[7]. The image dataset has a total of 44 classes. Of the 44 classes, there are 31 mammals, 9 birds, and 4 reptiles. Moreover, we couldn't find research-grade audio data for groups of animals other than birds. Thus, the audio model comprises 23 classes of birds all found in Nepal.

3.1 Image Data from iNaturalist

iNaturalist is an online social network of people sharing biodiversity information. It has a large size of data which includes images, audio, and video of large number of species found around the world. The data is collected by users who submit it to the site under the public domain, Creative Commons, or with all rights reserved. iNaturalist encourages its users to opt for more open licenses like Creative Commons for the research community. We collected a total of 38,442 images spanning 44 species all found in Nepal. As the data contained some anomalous images, we had to manually remove them. After the cleaning step, the data was preprocessed as per the Efficient Net Model for better training. Of the 38,442 images collected, 50 images from each class were used as test datasets. API provided by iNaturalist was used for downloading the data. The bar chart in Figure 1 shows the frequency of each class of image.

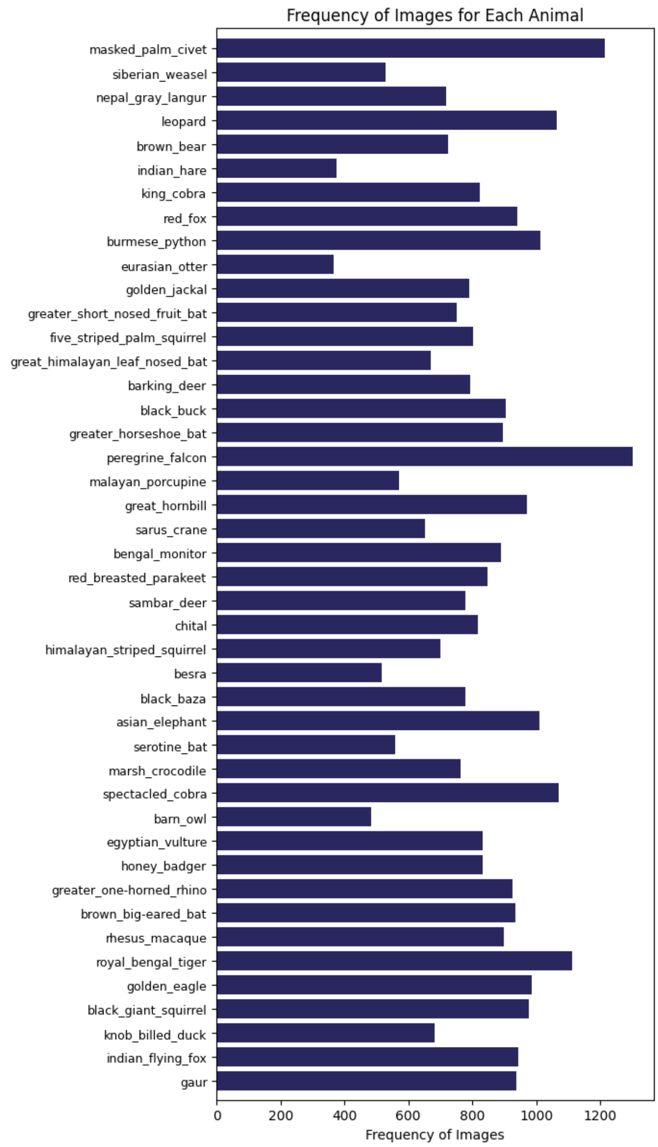

Figure 1: Image File count for each animal

3.2 Audio Data from Xeno-canto

Xeno-canto is a website dedicated to sharing wildlife sounds from all over the world. It provides its data under several licenses derived from Creative Commons licenses which are free for use for educational and research purposes. We collected a total of thousands of audio spanning 23 classes and performed temporal segmentation with each segment being 5 seconds long. A total of 16,230 audio files each of 5 sec was now our final dataset. Of the 16,230 audio files, 30 files for each class were used as the test dataset. API provided by Xeno-canto was used for downloading the data. The bar chart in Figure 2 shows the frequency of each class of audio.

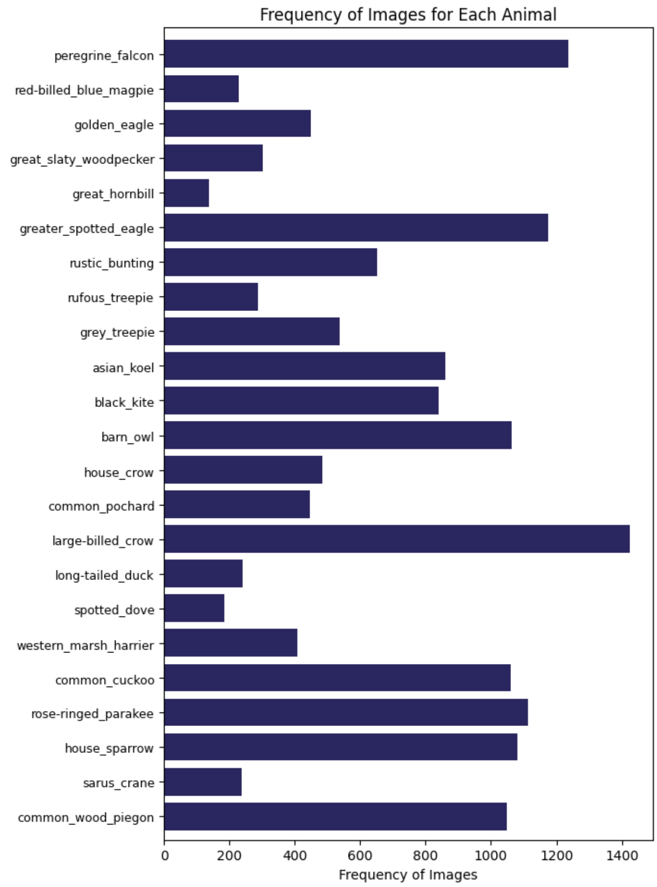

Figure 2: Audio File count for each bird

4. Proposed Methodology

4.1 Mel-frequency cepstrum Coefficients based on DCT-II for Audio Feature Extraction

Mel-frequency cepstrum (MFC) refers to the short-term power spectrum of a sound based on cosine transform. Mel-frequency cepstrum Coefficients(MFCCs) are coefficients that collectively make up an MFC. Mel-frequency cepstrum Coefficients transform is done on the audio to extract a small set of features that describes our audio. MFCC helps us capture the essential spectral characteristics of speech signals in a compact form, making them suitable for pattern recognition tasks.

```python
def mfcc_transform(y: np.ndarray, sr: int) -> torch.Tensor:
    melspec = librosa.feature.melspectrogram(
        y=y,
        sr=sr,
        n_fft=2048,
        win_length=2048,
        hop_length=512,
        n_mels=256,
        htk=True,
        norm="ortho",
    )

    mfcc_librosa = librosa.feature.mfcc(
        S=librosa.core.spectrum.power_to_db(melspec),
        n_mfcc=256,
        dct_type=2,
        norm="ortho"
    )

    return torch.from_numpy(mfcc_librosa)
```

Listing 1: MFCC Transformation Function

4.2 Expansion and Interpolation of Audio Data

After MFCC transformation, the tensors are transformed again to the proper dimension as required by EfficientNetB0 architecture. The following steps are performed:

- The input tensor y is replicated along one dimension to create three copies, preparing it for subsequent operations.
- The expanded tensor is resized to a fixed size of (224, 224) using bilinear interpolation, ensuring consistent dimensions for neural network input.
- Any extra dimensions introduced during resizing are removed, returning the tensor to its original shape.
- The processed tensor is returned for further use, completing the preprocessing steps commonly applied to image data for deep learning tasks.

```python
def expand_tensor_for_efficient_net(y) -> torch.Tensor:
    input_tensor_expanded = y.expand(3, -1, -1)
    resized_tensor =
        F.interpolate(input_tensor_expanded.unsqueeze(
        0), size=(224, 224), mode='bilinear',
            align_corners=False)

    resized_tensor = resized_tensor.squeeze()
    return resized_tensor
```

Listing 2: MFCC Transformation Function

4.3 Transformation as required by EfficientNetB0 Input Data

For best results and compatibility, we transformed our data as per the transformation required for EfficientNetB0 architecture.

```python
ImageClassification(
    crop_size=[224]
    resize_size=[256]
    mean=[0.485, 0.456, 0.406]
    std=[0.229, 0.224, 0.225]
    interpolation=InterpolationMode.BICUBIC
)
```

Listing 3: MFCC Transformation Function

4.4 Transfer Learning

Transfer learning is a technique in the field of machine learning where learned parameters on previously trained data are used to learn parameters of new data. This is done to reuse the learned parameters from previous data. Transfer learning is not only better for performance but also decreases the time required to learn features about subject in question.

4.5 EfficientNetB0

EfficientNet is a convolutional neural network architecture and scaling method that uniformly scales all dimensions of depth/width/resolution using a compound coefficient [8]. Unlike conventional practice that arbitrarily scales these factors, the EfficientNet scaling method uniformly scales network width, depth, and resolution with a set of fixed scaling coefficients.

4.6 Convolutional Neural Network

Convolutional Neural Networks are specialized kind of neural networks for processing data that has a grid-like topology [9]. Most machine learning libraries implement cross-correlation rather than conventionally known convolution operations in the realm of Digital Signal Processing. In conventional multi-layer perceptrons, large number of parameters are required to learn features where as CNN uses filters to stride through whole data using a lesser number of parameters aka parameter sharing.

5. Result and Analysis

Table 1: Overview of Results of Audio and Video Model

Model	Train Accuracy	Test Accuracy	F1 score on Test Data
Vision	96.00	86.10	86.12
Audio	98.36	90.39	88.12

5.1 Vision Model

The vision model was created with EfficientNetB0 as its base for transfer learning. We were able to achieve an accuracy of 86.12% with an F1-Score of 86.12 on the test dataset. Table 2 shows the various metrics of the vision model.

Table 2: Vision Model Metrics

Parameters	Metrics
Batch Size	32
Epochs	10
Pretrained Model	EfficientNetB0
Weights	ImageNet
Learning Rate	0.0001
Weight Decay	0.001
Loss Function	Categorical Cross Entropy
Optimizer	Adam
Total Parameters	4,063,912
Trainable Parameters	4,063,912
Non-Trainable Parameters	0

Figure 3 and Figure 4 show the accuracy and loss of vision model respectively. The confusion matrix for the vision model on the test dataset is shown in Figure 5.

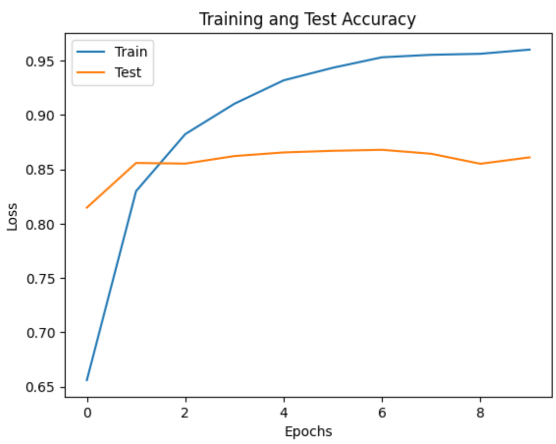

Figure 3: Accuracy of Vision Model on Train and Test Data

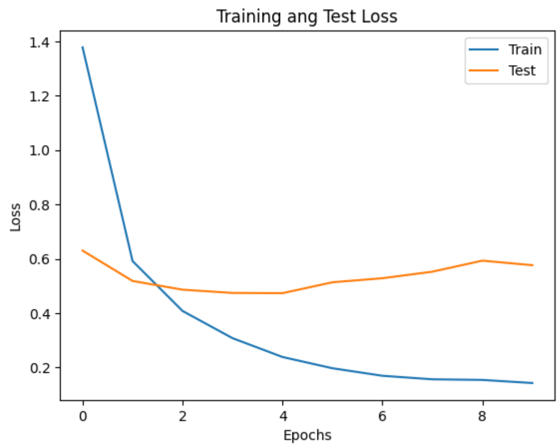

Figure 4: Loss of Vision Model on Train and Test Data

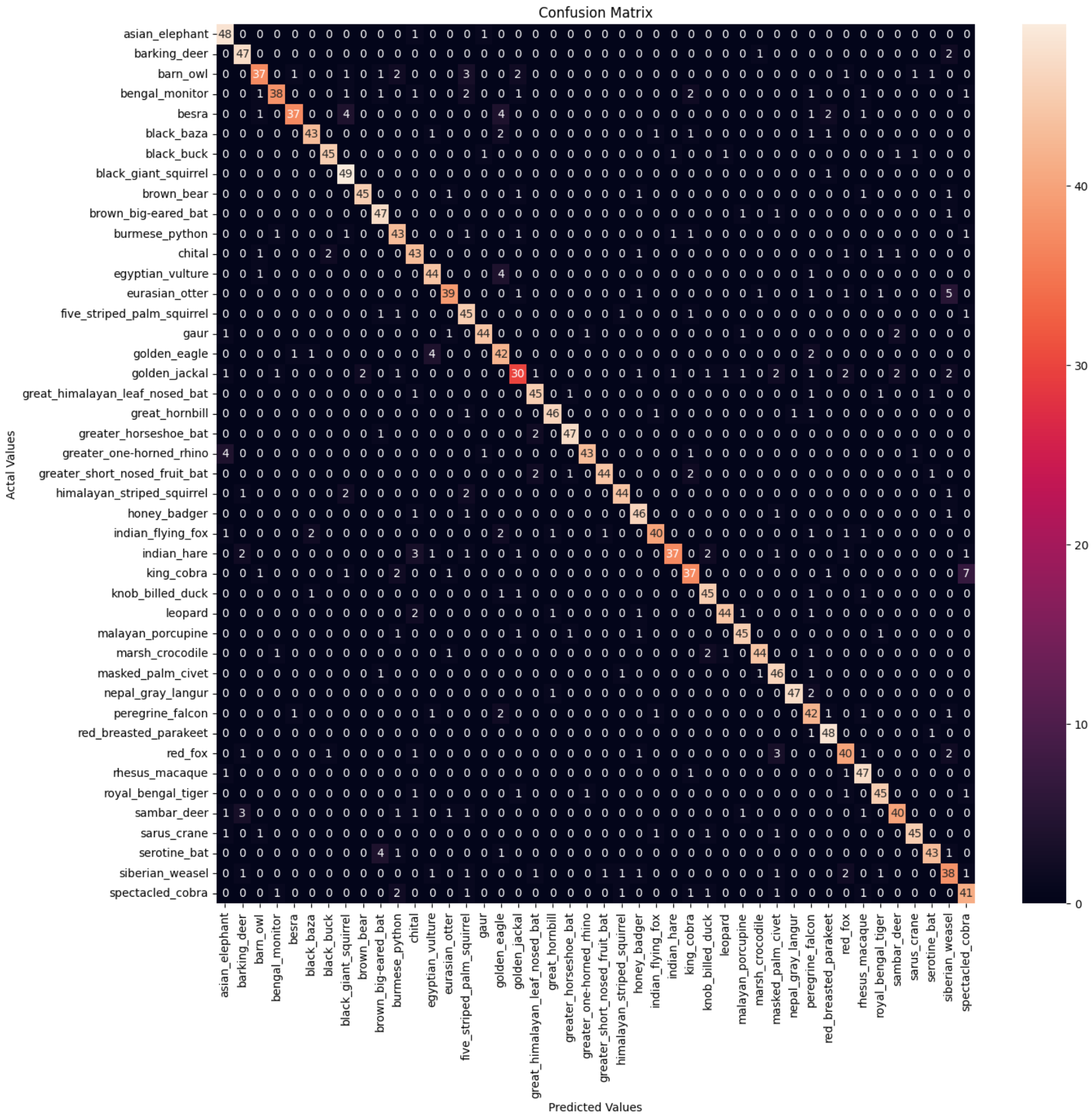

Figure 5: Confusion Matrix for Vision Model on Test Data

5.2 Audio Model

The audio model was also created with EfficientNetB0 as its base for transfer learning after feature extraction done with MFCC Transform which was again transformed as explained in Section 4.3 for input compatibility with EfficientNetB0.

We were able to achieve an accuracy of 90.39% and an F1-Score of 88.12 on the test dataset. The test size of each class is 30. Table 3 shows various metrics of the audio model.

We see a similar pattern as seen in the vision model. Here too, the audio model is struggling to classify audio of similar species. Figure 6 and Figure 7 show the accuracy and loss of the audio model respectively. Figure 8 shows the confusion matrix for the audio model on the test data.

Table 3: Audio Model Metrics

Parameters	Metrics
Batch Size	64
Epochs	15
Pretrained Model	EfficientNetB0
Weights	ImageNet
Learning Rate	0.0001
Weight Decay	0
Loss Function	Categorical Cross Entropy
Optimizer	Adam
Total Parameters	4,037,011
Trainable Parameters	4,037,011
Non-Trainable Parameters	0

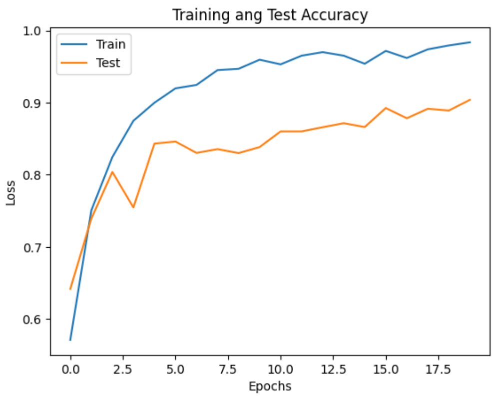

Figure 6: Accuracy of Audio Model on Train and Test Data

Figure 7: Loss of Audio Model on Train and Test Data

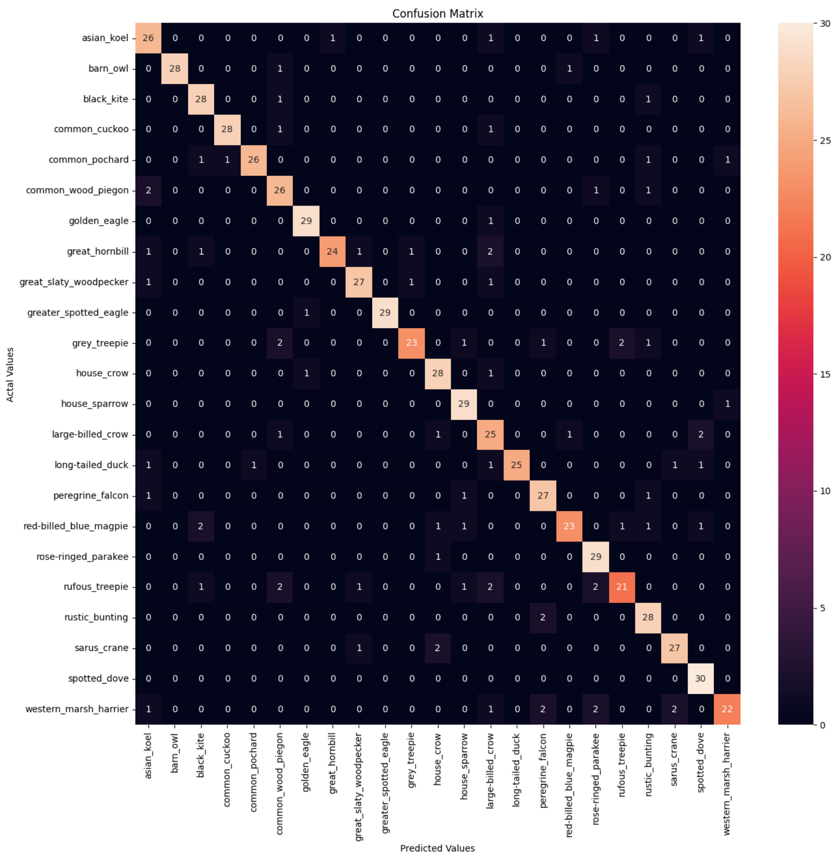

Figure 8: Confusion Matrix for Audio Model on Test Data

6. Conclusion

Our research work in wildlife classification combines audio and visual technologies, showing promising accuracy levels of 90.39% and 86.1% for audio and vision models on test data, respectively. Moreover, the audio and vision model achieves an F1 score of 86.12 and 88.12 on test data respectively. Despite challenges in distinguishing similar species, our study represents a significant advancement in leveraging technology for wildlife conservation. By enhancing monitoring capabilities, we aim to contribute to sustainable conservation practices and promote harmonious coexistence between humans and wildlife.

Acknowledgments

The authors would like to express their gratitude to everyone who supported them throughout this project. The authors are thankful for their aspiring guidance, invaluably constructive criticism, and friendly advice during the project work. The authors are grateful to them for sharing their views on issues related to the project. Without the aid of our friends and family members, our project would not have been attainable. This feat would not have been achieved without their assistance.

References

[1] Rajan Amin, Hem Sagar Baral, Babu Ram Lamichhane, Laxman Prasad Poudyal, Samantha Lee, Shant Raj Jnawali, Krishna Prasad Acharya, Gopal Prasad Upadhyaya, Megh Bahadur Pandey, Rinjan Shrestha, et al. The status of nepal's mammals. *Journal of Threatened Taxa*, 10(3):11361–11378, 2018.

[2] Tapil Prakash Rai, Sabin Adhikari, and Pablo Garcia Antón. An updated checklist of amphibians and reptiles of nepal. *ARCO-Nepal Newsletter*, 23:1–23, 2022.

[3] Hem Sagar Baral, Uba Raj Regmi, Laxman Prasad Poudyal, and Raju Acharya. Status and conservation of birds in nepal. *Biodiversity Conservation in Nepal: A Success Story. Department of National Parks and Wildlife Conservation, Kathmandu*, pages 71–100, 2012.

[4] Mohammad Sadegh Norouzzadeh, Anh Nguyen, Margaret Kosmala, Alexandra Swanson, Meredith S Palmer, Craig Packer, and Jeff Clune. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences*, 115(25):E5716–E5725, 2018.

[5] Rewan Gautam, Bhuwan Khatiwada, Bishwa Prakash Subedi, Niraj Duwal, and Kiran Chandra Dahal. Audio classifier for automatic identification of endangered bird species of nepal. 2023.

[6] Yan Qi, Cheng Baiyang, and Luo Lan. Deep learning based image recognition in animal husbandry. In *2021 18th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, pages 318–321. IEEE, 2021.

[7] Alexandra Sasha Luccioni and David Rolnick. Bugs in the data: How imagenet misrepresents biodiversity, 2022.

[8] Brett Koonce and Brett Koonce. Efficientnet. *Convolutional neural networks with swift for Tensorflow: image recognition and dataset categorization*, pages 109–123, 2021.

[9] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.