Detection of Bird Species Found in Bhutan Using Vision Transformer-based Transfer Learning

Karma Wangchuk[1*], Tandin Wangchuk[2], Tsheten Dorji[3]

Information Technology Department, College of Science and Technology, Royal University of Bhutan,
karma.cst@rub.edu.bt[1*], tandinwangchuk.cst@rub.edu.bt[2], tshetendorji.cst@rub.edu.bt[3]

Abstract

Birdwatching is an emerging recreational activity in Bhutan, attracting both local enthusiasts and international tourists due to the country's rich avian biodiversity. This growing interest contributes to local tourism and economic development. However, accurate bird identification remains a challenge due to variations in size, shape, and coloration, compounded by inconsistencies in English and Dzongkha nomenclature. Traditional identification methods, which rely on field guides and expert observations, are often prone to errors and disagreements. To address this limitation, we developed a bird detection and recognition system utilizing image processing and machine learning techniques. Bird images were collected from birdwatchers in Paro, Thimphu, and Trongsa, as well as from the Kaggle dataset. These images underwent preprocessing and augmentation to construct a comprehensive dataset. The study considered 23 bird species, and the model was fine-tuned using Google's pre-trained transformer encoder for image recognition, operating at a resolution of 244×244 with 16×16 patches. The model was trained on a dataset of 3,595 images, leading to a significant reduction in training and validation losses, from 2.8491 and 1.2231 to 0.0030 and 0.0529, respectively. The results indicate the effectiveness of the proposed approach in enhancing bird species identification, offering a valuable tool for birdwatchers and conservation efforts in Bhutan.

Key Words: *Vision Transformer, Bhutanese bird recognition, Transfer learning, Fine-tuning, Deep learning.*

1. INTRODUCTION

Bhutan's pristine forest cover provides a habitat for numerous species. According to Gyeltshen et al. (2020) the country is home to 748 bird species. Research on avian biodiversity has been gaining momentum, and the number of documented bird species is expected to increase in the coming years. In recent decades, ornithological studies have received growing attention, as research on bird behavior and population trends contributes to understanding environmental and climate change effects (Gavali & Banu, 2020; Gyeltshen et al., 2020). Moreover, the biological diversity and richness of Bhutan's largely undisturbed and unexplored environment attract birdwatchers from diverse backgrounds. Consequently, Bhutanese individuals engage in birdwatching either as a recreational activity or a supplementary source of income. Additionally, the tourism industry has experienced significant growth, increasing the demand for skilled professionals. However, the accurate identification of bird species remains a challenge. Ornithological expertise and skills are essential for the accurate identification of bird species. However, even professional birders occasionally disagree on species classification. Additionally, novice birdwatchers often find it challenging to distinguish between similar-looking birds, which can exceed the limits of human visual perception. While the human eye is adept at recognizing distinct objects with unique features, it is less effective in detecting minute details such as pixel variations. The application of Computer Vision techniques, specifically Image Processing, offers a more precise and reliable method for detecting and classifying bird species compared to human visual identification.

The traditional method of bird identification relies heavily on visual expertise and field guides. However, a method often proves insufficient with the diverse bird species found in Bhutan. Birders frequently encounter challenges while identifying birds from photographs, especially those taken from various angles and distances. Lack of an efficient and reliable identification system further exacerbates these challenges. This gap highlights the requirement for a system that can provide an accurate and consistent bird identification system to enhance the birdwatching experience and support the broader ornithological community.

2. RELATED WORK

The Convolutional Neural Network (CNN) is considered one of the best algorithms for Computer Vision and image content analysis

(Huang & Jeng, 2001). Krizhevsky et al. (2017) obtained state-of-the-art performance to classify the 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest dataset of 1000 different classes using deep learning. Variants of CNN such as LeNet, AlexNet (Krizhevsky et al., 2017), ZFNet (Zeiler & Fergus, 2014), GoogLeNet (Szegedy et al., 2015), ResNet (He et al., 2016), and VGGNet (Simonyan & Zisserman, 2015) algorithms have been used for image classification, detection, and recognition in computer vision applications. These algorithms won a series of annual ImageNet Large Scale Visual Recognition Challenge (ILSVRC) competitions.

The deep learning platform was developed to assist birdwatching in Taiwan using CNN (Huang & Basanta, 2019). They compared CNN, CNN with skip connection, and Support Vector Machine (SVM). CNN with skip connection obtained the best accuracy of 99.00%. Similarly, Niemi and Tanttu (2018) in Finland trained a bird identification system using CNN and deep learning algorithms. Furthermore, Ferreira et al. (2020), Stowell et al. (2019), and Sprengel et al. (2016) used deep CNN for bird detection and identification. Moreover, the modified You Only Look Once (YOLO) v4 model with a bird tracking algorithm based on a Kalman filter was implemented to identify a chicken's movement (Siriani et al., 2022). Similarly, Zou and Liang (2020) used the YOLO model to detect birds around transmission lines. Their model obtained a detection accuracy of 86.31%.

In this study, transfer learning was implemented. It is a powerful machine learning technique where a model developed for a particular task is reused as the starting point for a model on a second task. This approach is particularly beneficial in fields of image detection and recognition, where obtaining a large annotated dataset can be challenging and resource-intensive. In transfer learning, a pre-trained model on a large dataset is fine-tuned on a smaller, task-specific dataset. This technique allows for faster training and improved performance, especially when the target dataset is limited.

One of the seminal works in transfer learning for image recognition involves the use of CNNs. CNNs pre-trained on large datasets like ImageNet have been successfully adapted to a variety of tasks, achieving state-of-the-art results with minimal adjustments (Krizhevsky et al., 2017; Zeiler & Fergus, 2014). Research by Huh et al. (2016) demonstrated that features learned by deep CNNs are transferable and can be effectively used for a wide range of computer vision tasks, including object detection and image segmentation.

In the context of bird detection and recognition, transfer learning has been applied to fine-tune CNN models to identify bird species accurately. For instance, Niemi and Tanttu (2018) utilized transfer learning to train their bird identification system, achieving high accuracy by leveraging pre-trained models. Similarly, Huang and Basanta (2019) compared various CNN architectures, including those fine-tuned via transfer learning, to develop a robust birdwatching assistant in Taiwan, demonstrating the efficacy of this approach. Similarly, the transformer is used for object detection and classification.

Transformers, originally developed for natural language processing tasks, have recently been adapted for image recognition through architectures such as the Vision Transformer (ViT). ViTs apply the transformer model, which relies on self-attention mechanisms, directly to sequences of image patches, bypassing the need for convolutional layers. This novel approach has shown competitive performance compared to traditional CNNs.

Dosovitskiy et al. (2020) introduced the Vision Transformer (ViT), which pre-trains on large datasets and fine-tunes on specific tasks. The ViT model demonstrated that transformers could achieve state-of-the-art results in image classification tasks, given sufficient training data. Fine-tuning a pre-trained ViT on a smaller, domain-specific dataset allows the model to adapt to new tasks efficiently while retaining its powerful feature extraction capabilities.

The fine-tuning of transformer models for bird detection and recognition is an emerging area of research. Recent studies have begun examining the application of transformers to complex image recognition tasks. For instance, He et al. (2022) utilized fine-tuned transformer models to achieve high accuracy in species identification within ecological datasets. Their findings suggest that transformers have the potential to outperform traditional convolutional neural networks (CNNs), particularly in processing diverse and complex image data.

Integrating transfer learning with transformer fine-tuning presents a promising approach to improving image detection and recognition systems. By leveraging a pre-trained transformer

model and fine-tuning it on a specific dataset, researchers can utilize the model's extensive prior training and robust feature extraction capabilities while adapting it to the requirements of the target task.
This approach is particularly relevant in bird detection and recognition, where the ability to process varied and intricate image data is essential. Future research could explore the integration of Vision Transformers (ViTs) and other transformer-based models with transfer learning techniques to enhance the accuracy and efficiency of bird identification systems. These advancements have the potential to support ornithological studies and birdwatching while contributing to broader applications in wildlife monitoring and biodiversity conservation.

3. METHODOLOGY

The overview of the study is shown in Figure 1. There are six phases of this study. It begins with an extensive literature review, which is crucial to understand the current state of research in bird detection and recognition. Literature review involves examining existing studies on bird identification, machine learning algorithms, and image processing techniques. It also helps identify gaps in the current knowledge and inform the selection of appropriate methodologies and tools for developing the system. It ensures that the project builds on previous work and employs the most effective strategies to achieve its objectives. After the literature review, it is followed by data acquisition, data pre-processing, model selection, training, testing, and deployment. The main phases of the study are discussed in the following sub-section in detail.

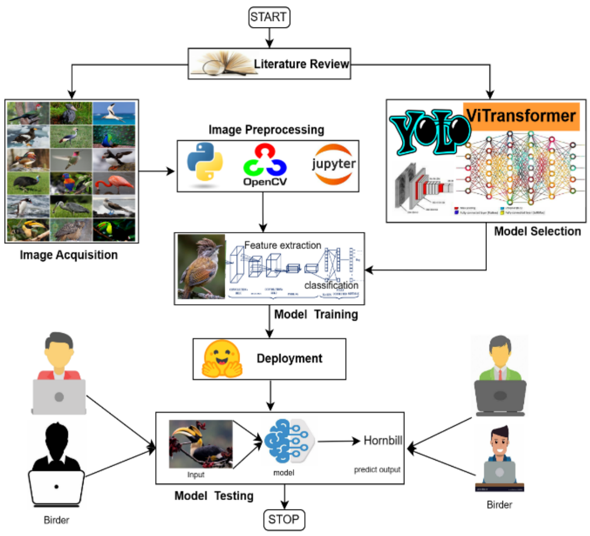

Fig. 1: Overview of the study.

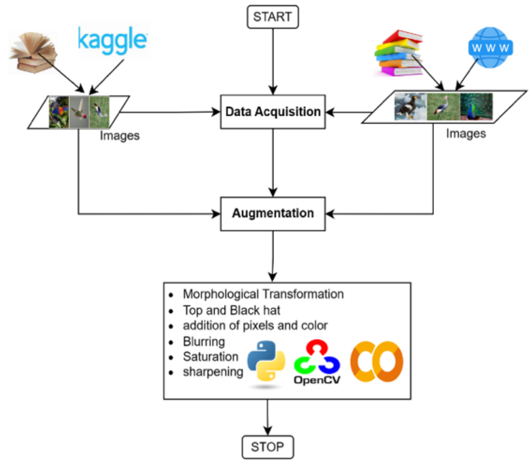

Fig. 2: Bhutanese bird dataset preparation pipeline.

3.1. Image acquisition and augmentation

Figure 2 illustrates the approach to compile and prepare a comprehensive dataset of avian species specifically found in Bhutan. The process begins with data acquisition, which involves sourcing images from multiple platforms to ensure a rich and diverse collection. The primary sources include online repositories Kaggle and birders from Paro, Thimphu, and Trongsa. The authenticity and accuracy of the species identified in these images are meticulously verified by cross-referencing with the authoritative resource "Birds of Bhutan: Habitat and Distribution" published in 2023. This step is crucial to maintain the integrity and scientific validity of the dataset. According to the Bhutan Birdlife Society (2023), number of bird species identified in Bhutan amounts to 767. However, a total of 23 birds were selected for the study and Dzongkha names for these birds were identified and confirmed.

Fig. 3: Augmented images using different techniques.

Furthermore, images were manipulated using several augmentation techniques to enhance the dataset's diversity and robustness. The morphological transformations, top and black hat transformations, and the addition of pixels and color variations were implemented using python and OpenCV. Moreover, techniques such as blurring, saturation adjustment, and sharpening were also applied as shown in Figure 3. These augmentations not only expand the dataset but also help in simulating various real-world conditions, thus making the dataset more comprehensive and suitable for training robust machine learning models.

Table 1: *English and Dzongkha names of curated image dataset.*

English	Dzongkha	Images
African Pied Hornbill	བྱ་དགོ་བོ	3187
Alphine Chough	ཅུང་ཀ	3166
Blue Heron	ཁྲུང་ཁྲུང་ཐལ་ཁ	3167
Canary	གསེར་བྱུ་རུ	3160
Crow	ཨོ་ལ	3163
European Turtle Dove	ཁི་ཁི་རྒྱལམོ	3138
Gila Woodpecker	ཏོག་ཏོ་ལ	3155
Golden Eagle	ཁྲ་གོད	3158
Himalayan Monal	བྱ་བབས / བྱ་མདངས	3151
Hoopoe	འདྲེ་ཏོག་ཟེམ	3155
House Martin	གསེར་ཁ	3157
House Sparrow	ནལ་གཟནམ	3155
Long-Eared Owl	ཙུག་པ	3160
Malachite Kingfisher	ཆུ་བྱ	3163
Mangrove Cuckoo	ཁུ་བྱུག་སྔོནམ	3137
Myna	ཏེབ་བྱ་གནགམོ	3141
Peacock	རྨ་བྱ	3156
Pigeon	ཕུ་རྒུ	3132
Red whiskered Bulbul	ཀི་ལི་སྔོར་སྔོར	3162
Snow Goose	ངང་པ	3190
Snowy Egret	ཆུ་བྱ་དཀརམོ	3132
Vulture	བྱ་གོད	3150
White Necked Raven	ཨོ་རོག	3160
Total		**72595**

Hundreds of images were randomly selected from each class for augmentation. This process produced 30 additional images for each original image, resulting in a total of 3,000 augmented images per class. After augmentation, the original images were combined with the augmented ones. The augmented images were compiled into a final dataset to be used in machine learning model training as shown in Table 1. The dataset of Bhutanese birds encompassed 23 distinct categories and contained a total of 72,595 images.

3.2. ViT model fine-tuning

The Vision Transformer (ViT) is a deep learning model designed for image classification tasks by Dosovitskiy et al. (2020) . ViT applies the principles of transformer architectures that was originally developed for natural language processing. Transformers were introduced in the seminal paper "*Attention is All You Need*" by Vaswani et al. in 2017 and revolutionized NLP by effectively modeling relationships within sequential data using self-attention mechanisms. This allowed models to process entire sequences simultaneously rather than sequentially, improving efficiency and scalability. Adapting these principles to vision tasks, the Vision Transformer aims to address the limitations of traditional convolutional neural networks (CNNs), which, despite their success, have constraints in modeling long-range dependencies and scalability.

The model was fine-tuned using the ViT model using the Hugging Face Transformers library. Furthermore, a pre-trained ViT model called *vit-base-patch16-224-in21k* from Google's library was implemented. The structure of the ViT begins with dividing the input image into fixed-size patches, such as 16x16 pixels as shown in Figure 4. Each patch is flattened and linearly embedded into a vector of a specified dimension. Positional embedding is then added to this patch embedding to retain spatial information. This sequence of patch embedding is fed into a standard transformer encoder, consisting of multiple layers of self-attention and feed-forward neural networks. Self-attention enables the model to weigh the importance of different patches when processing a given patch, allowing it to capture complex relationships across the entire image. Finally, the output of the transformer encoder is passed through a classification head, usually a fully connected layer, to produce the final classification logits for the image.

The model training was conducted using the ViT model using the Hugging Face Transformers library. Furthermore, a pre-trained ViT model called *vit-base-patch16-224-in21k* from Google's library was implemented.

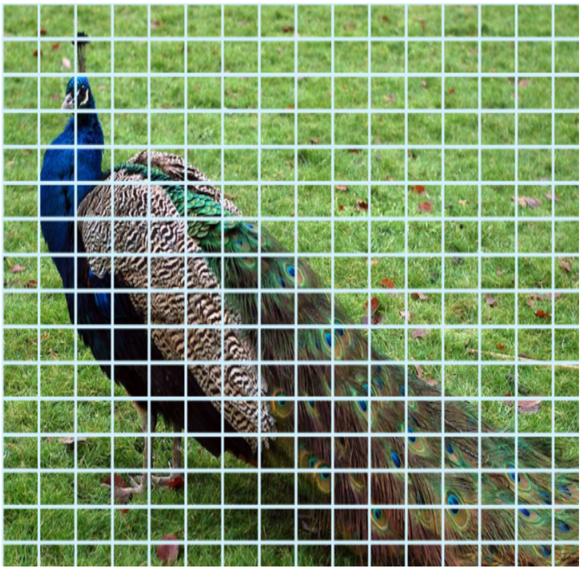

***Fig. 4:** 16x16 patches of an image as proposed by (Dosovitskiy et al., 2020)*

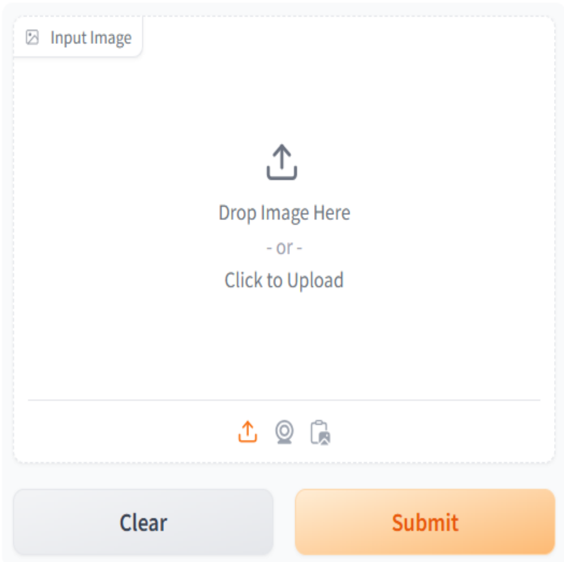

***Fig. 5:** Bird detection GUI on Hugging Face's Space.*

The model was deployed on Hugging Face's space. The GUI features three primary components: the input image section, the clear button, and the submit button. The input image section offers three key functionalities such as the option to drop an image or click to upload, access to the webcam, and the ability to paste an image from the clipboard. Users can load an image through any of these methods and submit it for prediction using the interface. If an incorrect image is uploaded, the clear button can be used to delete the uploaded image and reset the process. Once the correct image has been loaded, users can click the submit button to predict the name of the bird, as demonstrated in Figure 5.

***Table 2:** Image dataset used for model fine tuning.*

English	Dzongkha	Images
African Pied Hornbill	བྱ་དགོ་པོ	187
Alphine Chough	ཅུང་ཀ	166
Blue Heron	ཁྲུང་ཁྲུང་ཐལ་ཁ	167
Canary	གསེར་བྱ་ཙ	160
Crow	ཨོ་ལ	163
European Turtle Dove	ཐི་ཐི་རྒྱལམོ	138
Gila Woodpecker	ཏོག་ཏོ་ལ	155
Golden Eagle	ཁྲ་རྒོད	158
Himalayan Monal	བྱ་བབས / བྱ་མདངས	151
Hoopoe	འདྲེ་ཏོག་ཟེམ	155
House Martin	གསེར་ཁ	157
House Sparrow	ནལ་གཟནམ	155
Long-Eared Owl	ཨུག་པ	160
Malachite Kingfisher	ཆུ་བྱ	163
Mangrove Cuckoo	ཁུ་བྱུག་སྔོནམ	137
Myna	ཏེབ་བྱ་གནགམོ	141
Peacock	རྨ་བྱ	156
Pigeon	ཕུ་གུ	132
Red whiskered Bulbul	ཀི་ལི་སྔོར་སྔོར	162
Snow Goose	ངང་པ	190
Snowy Egret	ཆུ་བྱ་དཀརམོ	132
Vulture	བྱ་རྒོད	150
White Necked Raven	ཨོ་རོག	160
Total		**3595**

4. RESULT AND DISCUSSION

The model was fine-tuned using Google Colab. Google provides a free GPU. The following (Table 2) are the specifications of the system used for the model training: The model was fine-tuned using 1xTesla T4, 2560 CUDA core GPU. The Bhutanese birds' detection and recognition system was developed by fine-tuning Google's Vision Transformer (ViT) pre-trained model. The ViT works by breaking the input image into small, fixed-size 16 x 16 patch pixels. Each patch is flattened into a vector and given a positional embedding to retain spatial information. These vectors are fed into a transformer encoder, which uses self-attention to understand the relationships between different patches. Finally, the encoder's output is passed through a fully connected layer to produce the prediction of the bird.

The model was fine-tuned using 23 classes of birds comprising 3595 images with the African Pied Hornbill having the highest representation at 187 images, while the pigeon and snowy Egret had the lowest number of images, each totaling 132 as shown in Table 2. The model was trained for 4 epochs that run for 7639 steps with a batch size of 8. A step refers to a single training iteration where the model processes a batch of training data and updates its internal parameters based on the calculated loss. The number of steps per epoch depends on the batch size.

Figure 6 shows the training loss over a series of training steps, with the x-axis representing steps from 0 to 7000 and the y-axis showing the training loss. Initially, the loss is high, but it drops quickly during the first 1000 steps, indicating that the model is rapidly learning and adjusting its parameters to minimize errors. This steep decline signifies that the model is effectively capturing patterns in the training data.

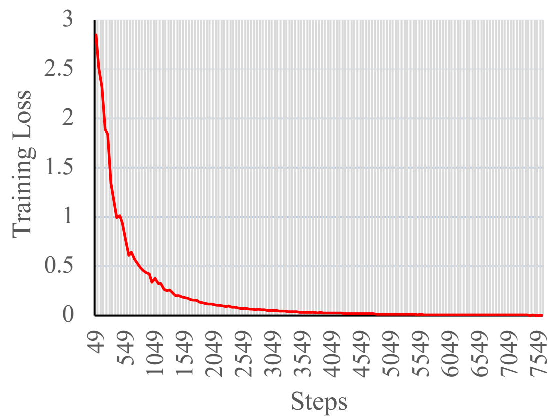

Fig. 6: *Model fine-tuning training loss.*

After the first 1000 steps, the loss continues to decrease but at a much slower rate. The model fine-tunes its parameters between 1000 and 3000 steps leading to incremental improvements in performance. Beyond 3000 steps, the loss curve flattens, indicating that the model is approaching its optimal performance. By 7000 steps, the loss is very low, close to zero, showing that the model has effectively minimized errors and learned well from the training data. The overall trend of rapid initial improvement followed by gradual convergence suggests a successful training process.

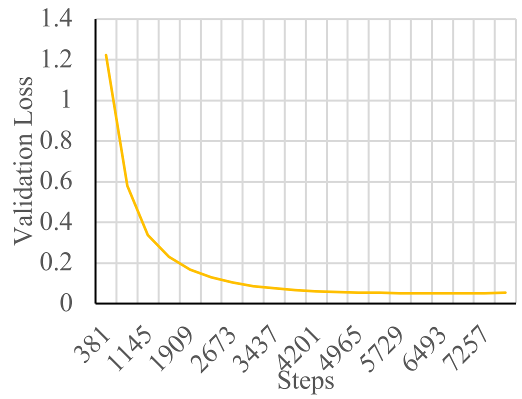

Fig. 7: *Model fine-tuning validation loss.*

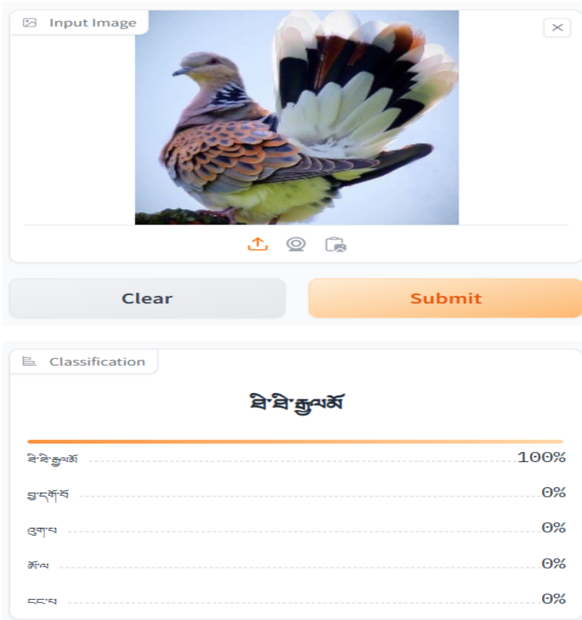

Fig. 8: *Dove prediction in Dzongkha.*

Similarly, Figure 7 illustrates the validation loss lover series of training steps. Initially, the validation loss was high. However, it plummets from 1.2 to 0.0529 after training for 4 epochs over several steps.

The model was deployed using Hugging Face's space and can be accessed from the link: *https://huggingface.co/spaces/KarmaCST/KarmaCST-Bhutanese_Bird*. Figures 8, 9, and 10 shows few sample outputs predicted by the Bhutanese bird detection and recognition system.

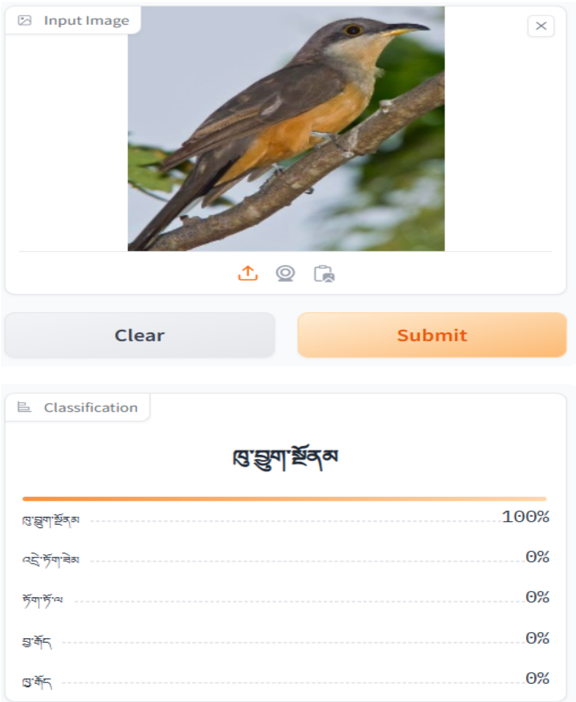

Fig. 9: *Cuckoo prediction in Dzongkha.*

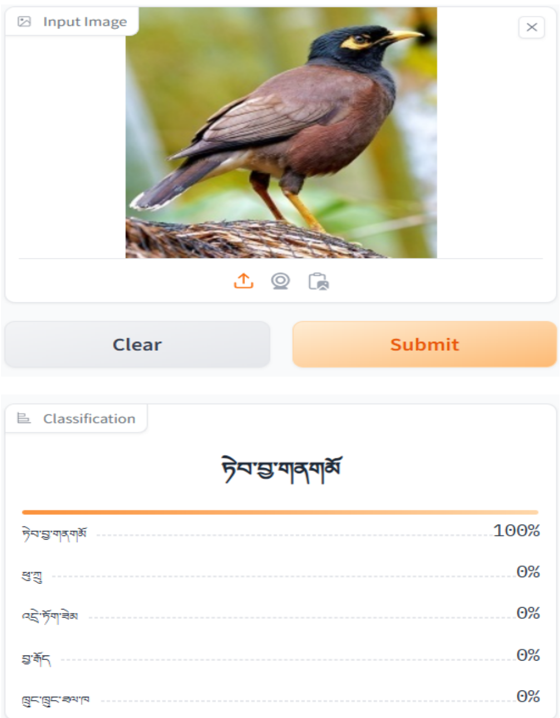

Fig. 10: *Myna prediction in Dzongkha.*

5. CONCLUSION

This study presents a Bhutanese bird detection and recognition system utilizing a machine learning approach. Bhutan is home to 748 bird species (Gyeltshen et al., 2020); however, acquiring images of these species remains challenging. Additionally, Dzongkha names for many bird species are not readily available. For this study, 23 bird species were selected, and their Dzongkha names were identified and verified using multiple sources. Bird images were collected from local birdwatchers and the Kaggle database. These images were then augmented to construct a comprehensive dataset. Two datasets were curated: one consisting of 72,595 images with augmentation and another with 3,595 images without augmentation.

The model was trained using Google Colab's free version and fine-tuned with Google's pre-trained Vision Transformer (ViT) model. ViT has demonstrated superior performance over Convolutional Neural Networks (CNNs) in computer vision tasks and was therefore selected for this study. The images, with a resolution of 224 × 244 pixels, were divided into 16 × 16 patches and subsequently vectorized. Each patch was vectorized, and positional embeddings were incorporated to retain spatial information. The patch embeddings were then processed by the transformer encoder, which consisted of multiple layers of self-attention mechanisms and feed-forward neural networks. The output from the transformer encoder was passed to the classification head to generate predictions.

The model was fine-tuned over four epochs, comprising approximately 7,000 training steps. In each step, the model processed a batch of data, computed gradients, and updated its weights. The training and validation losses decreased from 2.8491 and 1.2231 to 0.0030 and 0.0529, respectively, thereby optimizing predictive performance. The final model was deployed using Hugging Face's Spaces, allowing users to access and test the bird detection and recognition system online.

Future research could leverage this dataset to further enhance and refine the system. Additionally, the inclusion of more bird species and the exploration of alternative machine learning algorithms may contribute to improved accuracy. Moreover, developing a mobile application and a dedicated website could enhance accessibility and usability for a broader audience.

REFERENCES

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv preprint arXiv:2010.11929.

Ferreira, A. C., Silva, L. R., Renna, F., Brandl, H. B., Renoult, J. P., Farine, D. R., Covas, R., &

Doutrelant, C. (2020). Deep learning-based methods for individual recognition in small birds. *Methods in Ecology and Evolution*, 11(9), 1072–1085. https://doi.org/10.1111/2041-210X.13436

Gavali, P., & Banu, J. S. (2020). Bird Species Identification using Deep Learning on GPU platform. *International Conference on Emerging Trends in Information Technology and Engineering*, Ic-ETITE 2020. https://doi.org/10.1109/IC-ETITE47903.2020.85

Gyeltshen, P., Norbu, C., & Rinchen, K. (2020). Avifauna found in the State Reserved Forest Land of Trongsa district, Bhutan. Bhutan *Journal of Natural Resources and Development*, 7(1), 55–72. http://www.bjnrd.org/index.php/bjnrd/article/view/47

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 770–778. http://image-net.org/challenges/LSVRC/2015/

Huang, C.-L., & Jeng, S.-H. (2001). A model-based hand gesture recognition system. *Machine Vision and Applications*, 12, 243–258.

Huang, Y. P., & Basanta, H. (2019). Bird Image Retrieval and Recognition Using a Deep Learning Platform. *IEEE* Access, 66980–66989.

Huh, M., Agrawal, P., & Efros, A. A. (2016). What makes ImageNet good for transfer learning? arXiv preprint arXiv:1608.08614.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet Classification with Deep Convolutional Neural Networks. *Communications of the ACM*, 60(6), 84–90. https://doi.org/10.1145/3065386

Niemi, J., & Tanttu, J. T. (2018). Deep learning case study for automatic bird identification. *Applied Sciences*, 8(11).

Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale *Image Recognition. ArXiv* Preprint, 1409–1556. http://www.robots.ox.ac.uk/

Siriani, A. L. R., Kodaira, V., Mehdizadeh, S. A., de Alencar Nääs, I., de Moura, D. J., & Pereira, D. F. (2022). Detection and tracking of chickens in low-light images using YOLO network and Kalman filter. *Neural Computing and Applications,* 34(24), 21987–21997. https://doi.org/10.1007/S00521-022-07664-W

Sprengel, E., Jaggi, M., Kilcher, Y., & Hofmann, T. (2016). Audio based bird species identification using deep learning techniques. *Infoscience.* https://infoscience.epfl.ch/record/229232/files/16090547.pdf

Stowell, D., Wood, M. D., Pamuła, H., Stylianou, Y., & Glotin, H. (2019). Automatic acoustic detection of birds through deep learning: The first Bird Audio Detection challenge. *Methods in Ecology and Evolution*, 10(3), 368–380. https://doi.org/10.1111/2041-210X.13103

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 1–9. https://www.cv-foundation.org/openaccess/content_cvpr_2015/html/Szegedy_Going_Deeper_With_2015_CVPR_paper.html

Zeiler, M. D., & Fergus, R. (2014). Visualizing and Understanding Convolutional Networks. *Computer Vision–ECCV 2014: 13th European Conference*, 818–833.

Zou, C., & Liang, Y. quan. (2020). Bird Detection on Transmission Lines Based on DC-YOLO Model. *IFIP Advances in Information and Communication Technology, 581 AICT,* 222–231. https://doi.org/10.1007/978-3-030-46931-3_21