# Design Report: New Zealand Bird Sound Classifier

Harry Wills
*Victoria University of Wellington*
Wellington, New Zealand
willsharr@myvuw.ac.nz

***Abstract*—This project develops a machine learning model to classify New Zealand bird species from vocalisations, using audio data from the Xeno-canto database. The approach involves segmenting recordings, converting clips into Mel Spectrograms, and finetuning a Vision Transformer (ViT) model using Low-Rank Adaptation (LoRA). The model will initially be trained on a subset of species and progressively expanded to cover all 144 species, with real-world applications in biodiversity monitoring. Expected outcomes include high classification accuracy and a user-facing pipeline for field audio input, supporting scalable ecological conservation efforts.**

***Index Terms*—bird sound classification, machine learning, audio processing, biodiversity monitoring, vision transformer**

## I. Introduction

New Zealand's unique avian biodiversity is culturally and ecologically significant, but increasingly threatened by habitat loss and invasive species [1]. Manual identification of bird calls is time-consuming, creating a need for automated tools to aid conservation. This report presents a machine learning model that classifies New Zealand bird species by their vocalisations. The system segments and converts recordings into Mel Spectrograms, then finetunes a Vision Transformer (ViT) model using Low-Rank Adaptation (LoRA). The approach begins with a focused subset of species to establish baseline performance and is designed to scale to the full set of 144 species. This enables scalable, accurate bird sound classification across Aotearoa.

## II. Theory and Design

The project leverages audio signal processing and vision-based machine learning to classify bird sounds. Audio recordings are transformed into Mel Spectrograms, which represent frequency content over time, suitable for vision models like the Vision Transformer (ViT) [3]. ViT, originally designed for image classification, interprets the spectrograms as visual inputs, offering a novel approach to audio classification.

### *A. Data Collection and Preprocessing*

Audio data is sourced via the Xeno-canto API, providing 1,518 recordings of 144 New Zealand bird species with an average duration of 80.61 seconds. For initial development, we selected 10 species resulting in 554 recordings, including Tūī, Bellbird, Kākā, Fantail, Robin, Tomtit, Whitehead, Morepork, Saddleback, and Silvereye. These species provide a manageable baseline to test the pipeline. Audio files with a sample rate <22kHz are filtered out, with the rest being resampled to 44.1 kHz. Recordings are segmented into fixed-length clips (e.g., 5 seconds) and filtered using a quality threshold. Mel Spectrograms are generated using a 128 Mel filter bank for model input. The dataset will be progressively expanded to 25, then 50 species, and eventually $> 100$ as performance and scalability are validated.

### *B. Model Architecture and Finetuning*

The Vision Transformer (ViT) model is chosen for its ability to handle high-dimensional data like Mel Spectrograms. The architecture consists of multiple transformer encoder layers with self-attention, enabling temporal and frequency pattern extraction. A pretrained ViT is finetuned using Low-Rank Adaptation (LoRA), which modifies only a small subset of weights via low-rank decomposition [4], reducing training cost while maintaining accuracy. Users will be able to upload audio files, which are converted to spectrograms and classified by the model. The training strategy will be incremental, retraining or extending the model as new species are added.

## III. Expected Results

The finetuned ViT model is expected to achieve $\geq$85% classification accuracy on a held-out test set, with evaluation using precision, recall, and F1-score across all classes. Generalisation will be tested on recordings labelled as 'Unknown' in Xeno-canto to assess robustness to unseen species.

The system will initially be trained on a core set of 10 species to validate the architecture and pipeline. Once performance stabilises, we will expand the training set to include 25, then 50 species, ultimately targeting the full set of 144 native birds. Each expansion phase will involve iterative training, additional data augmentation, and balancing techniques to ensure performance is retained across a growing class set. This phased rollout supports a scalable and robust deployment.

Deliverables include a trained classifier, a spectrogram preprocessing tool, and an inference pipeline. These tools enable automated, field-deployable species identification to support ecological research and conservation efforts across Aotearoa.

## IV. Statement

All code and analysis in this project were completed by Harry Wills. Tools used include: Python, PyTorch, torchaudio, librosa, HuggingFace Transformers, and the Xeno-canto API. The full implementation and codebase are available at: https://github.com/harrywillss/NZ-Bird-Sound-Classifier.

# REFERENCES

[1] J. E. Dowding and E. C. Murphy, “The impact of predation by introduced mammals on endemic shorebirds in New Zealand,” Biological Conservation, vol. 99, pp. 47–64, 2001.

[2] R. S. Dhillon and B. M. Welling, “Audio event detection using Mel Spectrograms and convolutional neural networks,” Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process., pp. 1234–1238, 2019.

[3] A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” arXiv:2010.11929, 2020. [Online]. Available: https://arxiv.org/abs/2010.11929

[4] E. J. Hu et al., “LoRA: Low-rank adaptation of large language models,” arXiv:2106.09685, 2021. [Online]. Available: https://arxiv.org/abs/2106.09685