Data Collection and Dataset Creation

Data Sourcing:

- The PeaDetect dataset was assembled methodically by sourcing audio recordings from two primary repositories. Xeno-Canto and Freesound.
    - In search of samples for the target class, a query was conducted using its scientific name Pavo cristatus in Xeno Canto repository.
        - This search yielded 332 recordings of varying lengths.
        - **Positive samples** consist of different types of peafowl vocalizations (alarm calls, mating calls, territorial calls, etc.).
    - Negative samples, representing the absence of peafowl vocalizations, were selected to balance the binary classification dataset.
        - Species with **similar tonal and temporal qualities** were selected,particularly from the Sri Lankan ecological context.
        - Hence, a list of species such as fowls, coucals, eagles, parrots, owls, hornbills, and cranes were identified to represent negative samples.
        - Additionally, certain prominent ecoacoustic events exhibit vocal characteristics similar to peafowl calls.
    - For each of the identified species, relevant audio samples were queried and collected accordingly.
- **Filtering:**
    - After collecting all matching audio recordings, each sample was analysed for the filtering and annotation step.
    - All unsuitable audio samples were removed to ensure the data quality.
    - The recordings were then pre-processed to ensure uniformity in duration and quality before feature extraction and model training.
- **Segmentation:**
    - The audio recordings were segmented into 5-second intervals using Python libraries such as pydub and librosa.
    - After segmenting the audio, each segment was renamed following a consistent naming convention, incorporating the original sound identifier and a unique audio ID.
- **Annotation:**
    - A well-defined annotation methodology was essential for developing a high-quality dataset, facilitating the training of models capable of accurately detecting peafowl vocalizations.
    - Annotations were derived from existing dataset metadata and were manually listened and reviewed by the research team (human annotators) to ensure relevance and accuracy.
- **Dataset Creation:**
    - The dataset consists of two classes (presence and absence), where the absence class includes other bird species or environmental sounds.
    - The dataset is presented with the number of recording per class.
- **Final Dataset:**
    - The finalized PeaDetect dataset was subjected to a comprehensive analysis to assess its composition and quality, ensuring its suitability for subsequent modeling and research applications. The PeaDetect dataset comprises a total of 2950 audio samples, carefully divided into presence and absence categories. All audio samples have been standardized to a uniform duration and quality, adhering to a 44,100 Hz sample rate, 16-bit depth, and stereo channel configuration.