Aalto-yliopisto
Aalto-universitetet
Aalto University

Master's Programme in Automation and Electrical Engineering

Evaluation of Deep Learning Object Detection Architectures on Mel Spectrograms for Owl Acoustic Detection

Roland Vaatmann

Master's thesis
2025

Aalto-yliopisto
Aalto-universitetet
Aalto University

Author Roland Vaatmann

Title of thesis Evaluation of Deep Learning Object Detection Architectures on Mel Spectrograms for Owl Acoustic Detection

Programme Automation and Electrical Engineering

Major Control, Robotics and Autonomous Systems

Thesis supervisor D.Sc. (Tech) Joni Pajarinen

Thesis advisor(s) Ph.D Eetu Puttonen and Ph.D Hanna Rosti

Collaborative partner National Land Survey of Finland

Date 23.11.2025 **Number of pages** 67 **Language** English

Abstract

Reliable and scalable monitoring of Finnish owl populations faces urgent challenges as small species decline, and traditional surveys prove labour-intensive and often inconsistent. This thesis addresses these issues by evaluating whether Passive Acoustic Monitoring (PAM) combined with object-detection deep learning can provide accurate, fast, and operational detection of owl vocalisations, thereby supporting conservation and forest management decisions.

The research investigates the application of PAM and object-detection-based deep learning to detect three Finnish owl species: the Ural Owl (Strix uralensis), the Boreal Owl (Aegolius funereus), and the Eurasian Pygmy Owl (Glaucidium passerinum). Mel spectrograms are utilised, and two convolutional neural network architectures, Faster R-CNN with a ResNet-50 backbone and YOLOv11-l, are trained on annotated recordings from multiple deployment sites, focusing on species-specific vocalisations. Model performance is evaluated using metrics such as mean Average Precision, accuracy, F1-score, and inference latency.

Results demonstrate that YOLOv11-l significantly outperforms Faster R-CNN, achieving an mAP50:95 of 75.7% and an F1-score of 96.8% with substantially lower inference times. In contrast, Faster R-CNN records an mAP50:95 of 66.9% and an F1-score of 96.3%. This research culminates in the development of an end-to-end processing pipeline that leverages YOLOv11-l to analyse multi-site recordings and produce accurate detection summaries. These findings highlight the effectiveness of modern one-stage detectors for large-scale owl monitoring in Finnish boreal forests, thereby supporting data-driven conservation and management efforts.

Keywords Passive acoustic monitoring, bioacoustics, owl vocalisation detection, mel spectrogram, deep learning, object detection

Aalto-yliopisto
Aalto-universitetet
Aalto University

Tekijä Roland Vaatmann

Työn nimi Syväoppivien objektintunnistusarkkitehtuurien arviointi pöllöjen akustisessa tunnistuksessa Mel-spektrogrammien avulla.

Koulutusohjelma Automaatio ja sähkötekniikka.

Pääaine Säätötekniikka, robotiikka ja autonomiset järjestelmät

Vastuuopettaja/valvoja D.Sc. (Tech) Joni Pajarinen

Työn ohjaaja(t) Ph.D Eetu Puttonen and Ph.D Hanna Rosti

Yhteistyötaho Maanmittauslaitos

Päivämäärä 23.11.2025 **Sivumäärä** 67 **Kieli** Englanti

Tiivistelmä

Suomalaisten pöllökantojen luotettava ja laajamittainen seuranta on haastavaa. Pienikokoiset lajit ovat katoamassa ja perinteiset laskentamenetelmät ovat työvoimavaltaisia ja usein epäjohdonmukaista. Tässä työssä tutkitaan, voidaanko passiivisen akustisen seurannan (PAM) ja objektintunnistukseen perustuvan syväoppimisen avulla tuottaa tarkkaa, nopeaa ja operatiivisesti hyödynnettävää pöllöjen ääntelyn tunnistusta, joka tukee suojelu- ja metsänhoitopäätöksiä.

Tässä työssä tutkitaan passiivisen akustisen seurannan (PAM) ja objektintunnistukseen perustuvan syväoppimisen soveltamista kolmen Suomessa esiintyvän pöllölajin tunnistamiseen: viirupöllö (Strix uralensis), helmipöllö (Aegolius funereus) ja varpuspöllö (Glaucidium passerinum). Mel-spektrogrammeihin perustuen koulutettiin kaksi konvoluutiohermoverkkoarkkitehtuuria, Faster R-CNN ResNet-50-taustaverkolla ja YOLOv11-l, käyttäen useilta mittauspaikoilta kerättyjä luokiteltuja äänitteitä, joissa keskityttiin lajikohtaisiin ääntelyihin. Mallien suorituskykyä arvioitiin mittareilla, kuten mean Average Precision (mAP), tarkkuus (accuracy), F1-score ja prosessointiaika (inference).

Tulokset osoittivat, että YOLOv11-l suoriutui selvästi paremmin kuin Faster R-CNN, saavuttaen mAP50:95-arvon 75,7 % ja F1-arvon 96,8 % huomattavasti lyhyemmällä prosessointiajalla. Faster R-CNN:n mAP50:95-arvo oli vertailussa 66,9 %. Tutkimus johti käsittelyputken kehittämiseen, joka hyödyntää YOLOv11-l-mallia useilta mittauspaikoilta kerättyjen äänitteiden analysointiin ja havaintokoosteiden tuottamiseen. Tulokset korostavat modernien syväoppi tunnistus arkkitehtuurien soveltuvuutta laajamittaiseen pöllöseurantaan Suomessa ja tukevat tiedepohjaista suojelu- ja hoitotyötä.

Avainsanat Passiivinen akustinen seuranta, bioakustiikka, pöllöjen äänten tunnistus, mel-spektrogrammit, syväoppiminen, objektintunnistus

Table of contents

Preface and acknowledgements

I want to express my sincere gratitude to my thesis supervisor, Professor Joni Pajarinen, and thesis advisors, Professor Eetu Puttonen and Research Coordinator Dr Hanna Rosti, for their invaluable feedback and guidance throughout the development of my master's thesis. I also sincerely thank the National Land Survey of Finland for providing the essential computational resources and materials necessary for this research. Additionally, I am grateful to Huawei Tech4All's Natureguardian project for funding all the Audiomoth devices used in this study.

Otaniemi, 23 November 2025
Roland Vaatmann

Symbols and abbreviations

Symbols

t	time [s]
f	frequency [Hz]
B	signal bandwidth [Hz]
F_s	sampling frequency [Hz]
$x[n]$	discrete-time signal
$X[k]$	discrete Fourier Transform of $x[n]$
N	number of samples
m	Time frame index
k	frequency-bin index
L	hop length
$\omega[n]$	window length
$P[m,k]$	power spectrum at frame m, bin k
$S_{log}[m,k]$	log power spectrogram
$H_b[k]$	b-th Mel filterbank coefficient at bin k
$M[m,b]$	Mel filterbank energy at frame m, band b
X	input spectrogram
θ	trainable parameters

Abbreviations

Adam	Adaptive Moment Estimation
AP	Average Precision
CNN	Convolutional Neural Network
dB	Decibel
DFT	Discrete Fourier Transform
FFT	Fast Fourier Transform
FN	False Negative
FP	False Positive
FPN	Feature Pyramid Network
GRU	Gated Recurrent Units
HMM	Hidden Markov Models
IoU	Intersection over Union
LSTM	Long Short-Term Memory
mAP	Mean Average Precision
MFCC	Mel-frequency Cepstral Coefficients
NMS	Non-Maximum Suppression
PAM	Passive Acoustic Monitoring

PAN	Path Aggregation Network
RNN	Recurrent Neural Network
RoI	Region of Interest
SD	Standard Deviation
SGD	Stochastic Gradient Descent
STFT	Short-Time Fourier Transform
SVM	Support Vector Machine
TN	True Negative
TP	True Positive

List of Figures

List of Tables

1 Introduction

Birds act as sensitive indicators of environmental quality and ecosystem changes [1]. Their vocal activity fluctuates in response to habitat quality and disturbance, rendering sound a valuable proxy for ecological health. However, a feasible interpretation of vocalisations requires methods that reliably classify species on a large scale [2]. As biodiversity loss accelerates, monitoring must cover large geographic areas and long periods while remaining cost-effective [3]. In response, deep-learning techniques for bird species classification from audio have rapidly advanced and now support applications in bioacoustics, ecology, and conservation [4]. By converting continuous passive-acoustic recordings into species-level detections, these systems enable the tracking of population trends and migratory patterns, providing standardised indicators of ecosystem health at operational levels [2].

In the Finnish boreal forest ecosystem, owls serve as important indicators of environmental changes, with research highlighting declines in smaller species, such as the Boreal Owl (Aegolius funereus) and the Eurasian Pygmy Owl (Glaucidium passerinum) [5]. Additionally, while larger species, such as the Ural Owl (Strix uralensis), may exhibit different population dynamics, they are also affected by habitat fragmentation and the loss of mature forests. The decline of smaller owls is particularly concerning, as traditional population counts are often unreliable, underscoring the need for automated acoustic detection methods. By employing call-based individual recognition, the accuracy of abundance estimates can be enhanced, thereby improving data reliability and supporting more effective forest management and conservation strategies.

Natural soundscapes are complex, comprising many co-occurring species, overlapping acoustic sources, and even seasonal variability [6]. This includes seasonal variability, ambient noise, and anthropogenic sounds. Traditional survey methods, such as visual counts, point transects, and expert listening, can be time-consuming, costly, and intrusive. By contrast, Passive Acoustic Monitoring (PAM) enables the continuous, low-cost, and non-invasive sampling of soundscapes [7]. Additionally, PAM can be coupled with machine learning solutions, such as deep learning, which can, in turn, automate the identification of birds. In practice, the focus shifts from visual identification to analysing vocalisations, since many species produce distinctive acoustic signatures that can be detected even in visually obstructed environments [8].

A widely adopted representation for automated analysis is the Mel spectrogram [9]. It compresses audio into a time-frequency representation that preserves the discriminative structure of the raw waveform while reducing dimensionality. Mel spectrograms are well-suited for deep learning, offering a

compact and interpretable input that has often outperformed other representations [10].

For audio classification and detection on Mel spectrograms, pre-trained Convolutional Neural Network models (CNN) have consistently performed well. ImageNet-pretrained backbones, such as VGG16 [11] and ResNet [12], effectively transfer to spectrogram images, and pre-trained YOLO [13] detectors fine-tuned on acoustic data deliver strong localisation and labelling of vocal events. Leveraging transfer learning in this way enhances sample efficiency, stabilises optimisation, and generally results in higher accuracy than training from scratch, which is beneficial in bioacoustics, as labelled data are often limited [13]-[14].

Although there has been significant progress in bird-audio analysis using deep learning and transfer learning [14]-[17], current solutions are still not well-suited for Finnish owls in boreal soundscapes. Furthermore, most existing solutions are generalised models trained on diverse, non-Finnish datasets, rather than detectors specifically tailored to species and regional characteristics. Consequently, they more frequently misidentify species and exhibit lower accuracy in detecting owl calls. Furthermore, few studies have assessed the balance between accuracy and latency. Therefore, for the three most common owl species in Finland: the Ural owl (Strix uralensis), the Boreal owl (Aegolius funereus), and the Eurasian Pygmy owl (Glaucidium passerinum), there is no published, fine-tuned, high-accuracy detector validated end-to-end on Mel-spectrogram inputs.

The overarching aim of this thesis is to evaluate and operationalise deep learning for owl acoustic detection using mel spectrograms. Accordingly, two research questions are investigated: first, under identical training and evaluation conditions, what are the strengths and weaknesses of YOLOv11 and Faster R-CNN [18] with a ResNet-50 backbone, in detecting vocalisations of Strix uralensis, Aegolius funereus, and Glaucidium passerinum, as measured by detection accuracy, precision, recall, and inference latency; second, given the best-performing detector from the first part, how can it be implemented into a reproducible deployment pipeline that processes raw multi-site recordings and produces standardised detection summaries suitable for ecological mapping. This study focuses on these three species, compares the two detector types using mel-spectrogram inputs from curated, labelled recordings, and then deploys the top model on raw audio from multiple Finnish sites collected during February and March.

Chapter 2 reviews the theory and related literature on PAM, spectrogram representations, and deep learning for bioacoustics object detection. Chapter 3 details the methodology, including data acquisition, preprocessing,

spectrogram generation, and model configurations. Chapter 4 presents the results of Experiment 1, focusing on model performance, accuracy, and efficiency, and on deployment to raw multisite data in Experiment 2. Chapter 5 discusses the implications, limitations, and future directions. Finally, Chapter 6 summarises the conclusions of this thesis.

2 Literature Review and Theoretical Background

This chapter begins with a comprehensive review of relevant literature to establish the foundation for this study. The primary aim of this thesis is to evaluate two distinct deep learning architectures for object detection and to develop an end-to-end pipeline integrating these models. Following this overview, the chapter introduces the core theoretical concepts that underpin the research. First, the discussion focuses on object audio detection, highlighting its significance within the context of the thesis. Next, convolutional neural networks (CNNs) are introduced, followed by a detailed discussion of object detection based on CNN architectures. Finally, the chapter concludes with an examination of signal processing techniques relevant to detection tasks.

2.1 Relevant Literature

Monitoring avian populations through acoustic methods has been widely recognised as a valuable approach for ecological research and conservation efforts. Historically, this methodology depended on expert observers employing visual detection or auditory identification to census bird species. Passive Acoustic Monitoring (PAM) has emerged as an increasingly valuable tool in ecological studies [7], offering a scalable and non-invasive means of surveying soundscapes over extensive spatial and temporal scales. Deploying autonomous recording units enables continuous capturing of soundscapes over large areas and extended periods [15]-[16]. Analysing the extensive volume of audio data for bird species identification presents a significant challenge. Initial computational approaches employed traditional pattern recognition techniques to address this challenge [18]-[19]. Later pattern recognition techniques for automated avian identification encompass approaches, such as frame-based methodologies utilising Hidden Markov Models (HMMs), to model sequences of acoustic features, as well as event-based techniques that depend on handcrafted measurements of detected calls [21]. These conventional systems typically utilise features, such as Mel-frequency Cepstral Coefficients (MFCCs), or other spectral descriptors as input variables [22]. For example, Zhao et al. [21] combined an energy-based detector with MFCC-derived autoregressive features and a Support Vector Machine (SVM) classifier to identify bird species in field recordings. Many of these traditional acoustic classification pipelines demonstrated feasibility through automation. However, they frequently encountered challenges across diverse contexts, such as noise interference, overlapping sound sources, and limited applicability [6].

A crucial aspect of automated bird identification involves the representation of audio data. Early systems favoured compact features, such as MFCCs,

which summarise the spectral envelope of short sound frames. MFCCs proved effective for simpler classification models and remain feasible in some instances. However, they discard much temporal and spectral detail [23]. A widely adopted alternative has been the Mel spectrogram [9], [24]. Mel spectrograms provide a two-dimensional representation of audio waveforms, capturing both spectral and temporal information. Empirical research indicates that employing Mel spectrograms as input features often results in higher classification accuracy compared to MFCCs across various applications. For instance, a fully convolutional network achieved 88% accuracy using Mel spectrograms and 79% accuracy with MFCCs on a multi-species bird dataset [25]. This can be explained as Mel spectrograms generally preserve important frequency-time patterns that may serve as distinctive features across different species. Nonetheless, MFCCs and other feature descriptors can provide complementary information. Prior research has demonstrated the efficacy of combining various feature types, such as in an experiment employing a dual-input CNN architecture that utilised both Mel spectrograms and MFCCs as inputs [26]. By doing this, they achieved an average precision of 97.5%, compared to 94-96% when using either representation alone. Overall, the field has shifted towards spectrogram-based representations, as they provide a more expressive and informative substrate for machine learning applications, particularly in the realm of deep learning.

Over the last ten years, deep learning has become an essential instrument in the field of automated bird identification. Deep learning architectures, notably Convolutional Neural Networks (CNNs), have increasingly outperformed traditional signal processing techniques in the domain of acoustic event detection and classification. This shift highlights the potential of advanced neural network models to enhance accuracy and reliability in complex audio analysis tasks [16]. Instead of depending on manually crafted features, CNNs can automatically learn features from spectrogram inputs. Early uses of CNNs in bird audio showed significant improvements in accuracy and efficiency. For instance, a study on deep networks achieved much higher retrieval rates than traditional methods for detecting bird calls in various recordings [27]. Another study in 2019 observed that modern machine learning can achieve high detection performance in remote-monitoring environments [28]. As dataset sizes grew, researchers increasingly utilised transfer-learning techniques [29]. Specifically, CNN architectures pre-trained on extensive image datasets, such as ImageNet [30], were subsequently finetuned on spectrogram data. This approach improved sample efficiency and performance, particularly when labelled audio data was scarce [26]-[27]. For instance, the authors adapted a pre-trained DenseNet from ImageNet to spectrograms, yielding better bird detection results than training from scratch [31]. Similarly, commonly used image models, such as VGG16, [11] and ResNet-50 [12] have been successfully adopted for bird audio spectrogram classification

[32]-[34]. Another larger state-of-the-art system, BirdNET, can identify 984 bird species by sound [16]. It is a ResNet-based model that is trained on crowdsourced datasets.

Beyond CNNs, researchers have also explored recurrent neural networks (RNNs) [35] and hybrid models to leverage temporal context in bird audio. Bird calls and songs unfold over time, making sequence modelling a beneficial approach. Two prominent variants of RNNs employed in this domain are Long Short-Term Memory (LSTM) [36] and Gated Recurrent Units (GRU) [37]. For example, the authors proposed an attention-based bidirectional GRU model for multi-species classification [38]. Their approach utilised sliding windows of MFCC features and an attention mechanism to focus on the most important frames. By aggregating predictions over time and assuming the highest-probability species as the dominant species per clip, their Bi-GRU classifier achieved an F1-score of 0.84 on a multi-species Xeno-Canto dataset [38].

Additionally, instead of outputting the presence of a species, detection algorithms have emerged that aim to localise each species' vocalisation in time within a recording. In machine learning, these are object detection CNNs. These object detection CNNs can directly output bounding boxes around the calls on a spectrogram image. For example, YOLO [13] and Faster R-CNN [18] were initially developed for image object detection, but they have been modified for acoustic data in many studies [27], [39]. The one-stage architecture of YOLO proved both efficient and effective for spotting calls in noise recordings [27]. In general, one-stage detectors offer fast inference suitable for real-time or low-latency monitoring, whereas two-stage detectors, such as Faster R-CNN, often yield high accuracy by first generating region proposals and then classifying them. Both approaches are used in bioacoustics. For instance, Li et al. [40] used Faster R-CNN to detect bird calls in continuous audio, finding it robust but computationally heavy for large datasets. The downside of bounding box-based models is that training such models demands detailed annotations which are labour-intensive to produce. Current research often compares these approaches or combines them.

2.2 Audio Object Detection with Convolutional Neural Networks

CNNs are deep learning architectures characterised by their use of local connectivity and weight-sharing mechanisms [13], [18]. When an audio signal is transformed into a time-frequency representation, such as a Mel spectrogram, CNNs are capable of learning hierarchical features, including edges, harmonic structures, and amplitude modulations, which are indicative of vocalisation patterns [18]. Formally, the network models a nonlinear mapping function $F(X, \theta)$, where X denotes the input spectrogram and θ represents the set of parameters optimised during training on labelled datasets. This image-centric inductive bias aligns naturally with the two-dimensional grid structure of the spectrogram which compactly encapsulates acoustic energy [14].

2.3 CNN-based Object Detection Architectures

Object detection frameworks extend CNN architectures by facilitating both the classification of target acoustic events and their precise spatial localisation within spectrogram representations [13], [18]. When applied to spectrogram analysis, these detectors generate time-frequency labels that delineate the onset and offset times, as well as the frequency extents of the acoustic phenomena, accompanied by class probability estimates [14], [41]. This approach enables detailed characterisation of acoustic events, advancing the capabilities of automated sound recognition systems [41].

Two common architectures for object detection are two-stage detectors, such as Faster R-CNN [18], which first generates class-agnostic region proposals, and then refines and classifies them. In contrast, one-stage detectors, such as the YOLO family, [13] perform dense predictions of classes and bounding boxes across multiple scales in a single forward pass. Both methods use features extracted by a convolutional neural network backbone and employ multi-scale feature representations. This is especially beneficial considering the significant variability in the duration and spectral bandwidth of acoustic events [7].

In the simplified schematic presented in Figure 1, Faster R-CNN receives a power-scaled Mel spectrogram as an input. The convolutional backbone is responsible for extracting multi-scale features. Initially, a region proposal network generates candidate regions which are then refined by a Region of Interest (RoI) head to produce final class predictions and localised event detections [18].

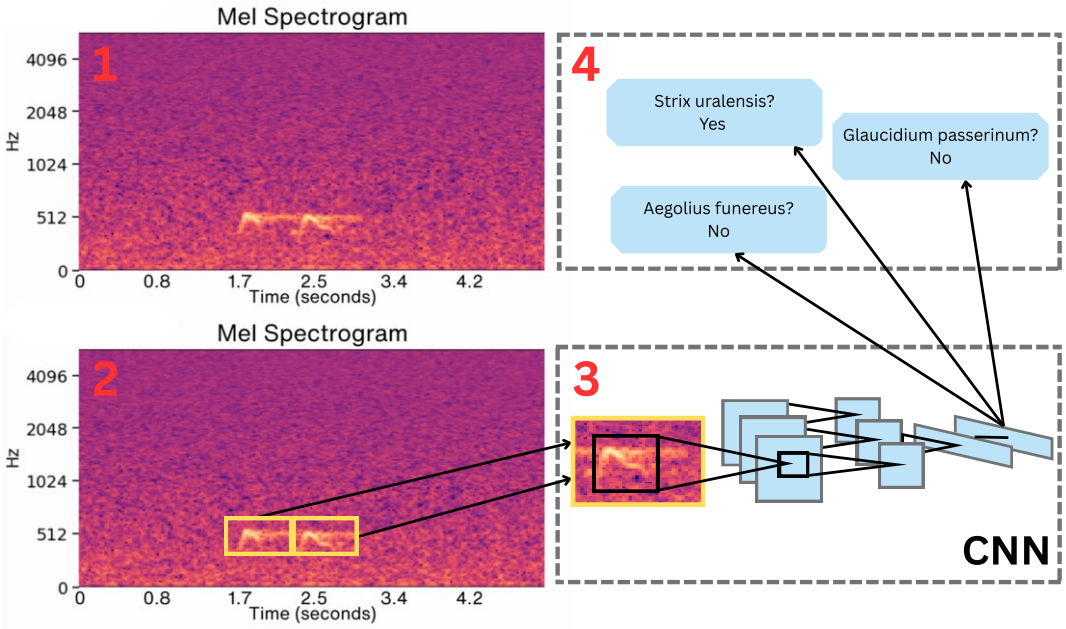

Figure 1: Overview of the simplified Faster R-CNN detection pipeline. The process begins with (1) a power-scaled Mel spectrogram serving as the input image. Next, (2) region proposals are generated to identify candidate vocalisations. Subsequently, (3) convolutional feature extraction is performed on each proposed region. Finally, (4) regions are classified to determine the specific target owl species.

2.3.1 Faster R-CNN

The Faster R-CNN framework is a prominent two-stage object detection methodology [18]. Its initial stage involves generating a set of candidate regions of interest which are subsequently classified and refined in the subsequent stage. The model employs a convolutional backbone network, such as ResNet-50 [12], to extract hierarchical feature representations from the input images, enhanced by a Feature Pyramid Network (FPN) [42] that facilitates multi-scale detection capabilities. This architectural design enables effective identification of objects across a broad spectrum of sizes, including both small and large instances.

In the first stage, known as the Region Proposal Network (RPN), a fully convolutional module processes the features from the backbone and FPN [18]. This module predicts objectness scores and bounding box regressions relative to a set of predefined anchors which vary in scale and aspect ratio. The most promising proposals are then selected through a Non-Maximum Suppression (NMS) procedure [43].

The second stage involves a detection head in which the proposed regions are aligned to fixed-size feature tensors using RoI Align. These regions are then processed by a per-region classifier and bounding box regressor which output class probabilities and refined bounding boxes. The training employs a multi-task loss function that combines classification accuracy with bounding box regression, utilising metrics, such as Smooth-L1 loss, and Intersection over Union (IoU) based methods [18]. During training, proposals are classified as positives or negatives based on IoU thresholds, with NMS applied to suppress redundant detections.

In this study, ResNet-50 serves as the backbone network in conjunction with the FPN, a choice motivated by its balance between detection accuracy and computational efficiency. Deeper residual networks have been shown to improve proposal quality and overall detection performance, particularly at higher IoU thresholds, consistent with findings reported in the literature on Faster R-CNN [12], [18].

2.3.2 YOLO11

The YOLO family constitutes a prominent approach within the domain of one-stage object detection methodologies [13]. Unlike traditional two-stage methods, YOLO systems facilitate a direct prediction of object categories and bounding box coordinates from dense feature maps through a single forward pass, thereby significantly enhancing computational efficiency [13]. The architecture typically comprises a convolutional backbone responsible for hierarchical feature extraction from the input image. This backbone is often augmented by an intermediary neck component, such as an FPN or a Path Aggregation Network (PAN) [13], which integrates multiscale information to enhance detection performance across a range of object sizes.

Subsequently, detection heads perform dense, per-location predictions, including bounding-box parameters, objectness scores, and class probabilities. This structural design enables the effective identification of objects present within a broad size spectrum within a single image frame.

The training process is guided by a comprehensive loss function, typically composed of three key components: an objectness term that distinguishes between foreground and background regions, a classification loss that assigns target categories, and a bounding-box regression loss based on IoU-related metrics. Modern variants of YOLO incorporate sophisticated techniques, such as dynamic label assignment and, in anchor-free configurations, predict offsets relative to grid-cell centres instead of relying on predefined anchor boxes [44]. During inference, non-maximum suppression, along with its soft counterpart, soft-NMS, is employed across multi-scale predictions to

eliminate redundant detections while retaining those with high confidence scores.
This study employed the YOLOv11 algorithm, which has demonstrated low latency and high accuracy in previous research [41]. The choice of this model is justified by its superior performance metrics which enhance real-time detection capabilities while maintaining reliable accuracy [45].

2.3.3 Optimizer

The optimizer determines the way that the parameters of a neural network are updated during training, with the objective of minimizing the loss function over time. After each mini-batch, it employs the computed gradients to guide the direction and magnitude of parameter updates within the parameter space.

Stochastic Gradient Descent (SGD) updates parameters by moving them in the direction opposite to the gradient of the loss function with respect to those parameters [46]. While this approach is straightforward and computationally efficient, it can introduce noise into the update process. Incorporating momentum into SGD smooths these updates by maintaining a running average of recent gradients, thereby facilitating navigation of flat or curved loss surfaces and accelerating convergence, provided the learning rate is calibrated correctly.

Adaptive Moment Estimation (Adam) extends the concept of SGD with momentum by maintaining two moving averages: one of the gradients, representing their overall direction, and one of their squared values, indicating their scale [47]. This adaptive approach allows for per-parameter adjustment of the learning rate, often resulting in easier hyperparameter tuning compared to standard SGD and typically yielding faster convergence, particularly in scenarios involving small or noisy batches [48].

AdamW is an optimisation algorithm that extends Adam by introducing a more principled approach to weight decay, commonly understood as parameter shrinkage [49]. Unlike the original Adam formulation, which integrates weight decay directly into the gradient update, AdamW decouples the weight decay term from the gradient-based optimisation step. This separation ensures a consistent and theoretically sound application of weight decay, leading to improved generalisation performance [49]. Furthermore, this decoupling enhances the interpretability of the weight decay hyperparameter. In practice, weight decay is typically applied to the weights of convolutional and linear layers, while it is generally not applied to bias parameters or normalisation layers.

2.4 Signal Processing

This section provides an overview of the process by which raw audio signals are transformed into features suitable for input into CNNs. Initially, temporal processing is conducted, which involves downsampling with anti-aliasing filtering and segmenting the audio into five-second chunks. Subsequently, Short-Time Fourier Transform (STFT) is applied to shift the data from the time domain to the frequency domain. During the frequency processing stage, a power spectrogram is generated and then processed through a Mel filterbank, which applies perceptual frequency warping. Logarithmic compression is subsequently applied to produce log-Mel spectrograms. The following subsections elaborate on these steps, covering spectral feature extraction, signal acquisition and sampling, Fast Fourier Transform (FFT) and STFT implementation, and the formation of power, log-power, and Mel representations. Signal processing steps in the time domain and frequency domain are illustrated in Figure 2.

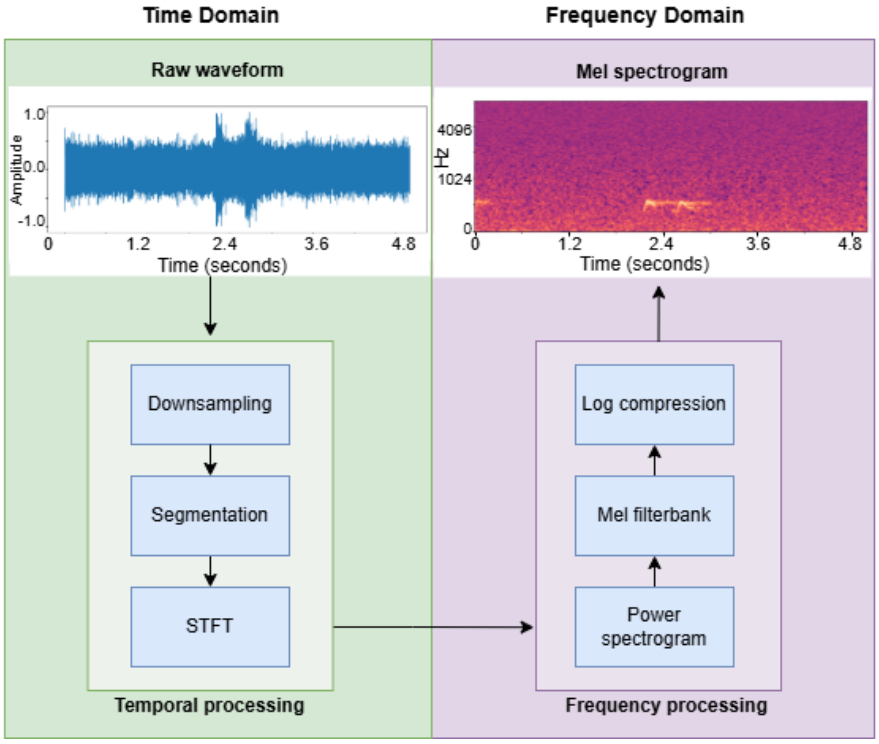

Figure 2: Signal processing steps in the time-domain and frequency-domain

2.4.1 Spectral Features

Raw audio recordings are inherently complex and often contain substantial noise, rendering it challenging to extract meaningful patterns [6]. Factors, such as environmental interference, variations in microphone characteristics, distance from the source, and overlapping vocalisations, further complicate analysis. Typically, the initial step involves representing the audio signal as amplitude over time, as depicted in Figure 3, which shows the vocalisation of Strix uralensis. While this time-domain representation captures envelope and intensity fluctuations, it does not reveal the spectral structure of the signal. To address this, a log-power spectrogram of the same recording (see Figure 4) provides a time-frequency representation that more clearly differentiates the Strix uralensis call from background noise, highlighting frequency bands and temporal modulations that are not visible in the waveform.

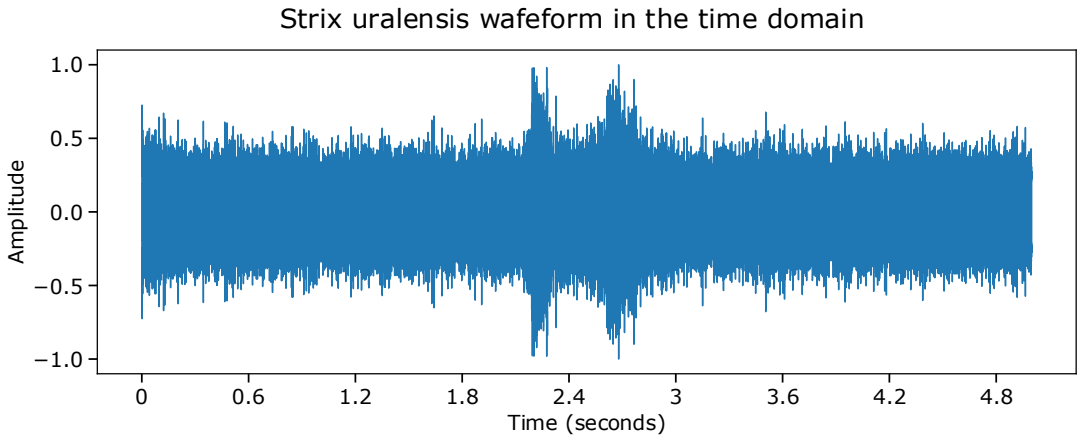

Figure 3: Linear amplitude over time of Strix Uralensis call

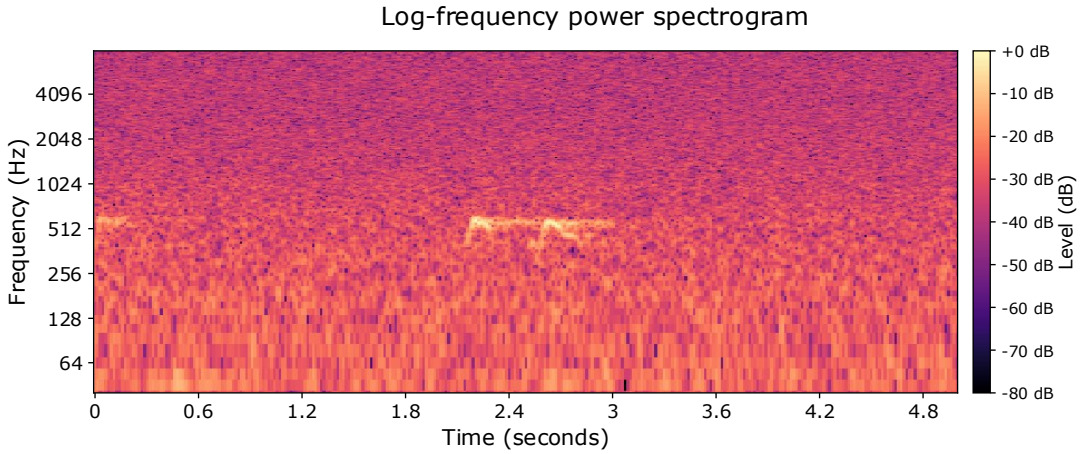

Figure 4: Log power spectrogram of raw waveform from Figure 3

2.4.2 Sampling the signal

Sound is typically captured by microphones capable of recording frequencies within the human audible range which ranges from approximately 20 Hz to 20,000 Hz. To accurately represent and reconstruct a recorded sound signal, the sampling rate must fulfil the Nyquist-Shannon sampling theorem which states that the sampling frequency must be at least twice the highest frequency component present in the original signal [50].

$$F_s > 2B \tag{2.1}$$

Here, B is the bandwidth of the signal, and F_s is the sampling rate. For example, a sampling rate of 20,000 Hz allows accurate reconstruction of frequency components up to 10,000 Hz.

2.4.3 Fourier Transform and Short-Time Fourier Transform

Analysing audio signals is commonly performed in the frequency domain, in which information about the frequencies present in a signal is extracted. The Fourier transform is a mathematical tool that enables this frequency analysis. It converts a time-domain signal into its constituent frequency components. The Discrete Fourier Transform (DFT) [51] is mathematically defined for a discrete-time signal $x[n]$ as follows:

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-\frac{i2\pi kn}{N}} \tag{2.2}$$

Here, $X[k]$ represents the frequency-domain representation of the time-domain signal x[n], and N is the number of samples.

However, audio signals, such as bird vocalizations, are typically non-stationary, meaning that their frequency characteristics vary significantly over time [52]. The standard Fourier transform alone cannot capture this dynamic behaviour. To overcome this limitation, the Short-Time Fourier Transform (STFT) is employed [34], [51]. The STFT divides the input signal into short overlapping segments, applying the Fourier transform to each segment separately, effectively producing frequency information localized in time. The mathematical definition of the discrete STFT is given by:

$$STFT\{x[n]\}(m,k) = \sum_{n=-\infty}^{\infty} x[n]\omega[n - mL] e^{-\frac{i2\pi kn}{N}} \tag{2.3}$$

Here, $x[n]$ is the input signal, $\omega[n]$ is a window length, m represents the segment index along the time axis, L is the hop length, and k denotes the frequency bin index. STFT provides a two-dimensional representation of the audio signal, with frequency components shown as they change over time. This characteristic renders it especially valuable for creating spectrograms.

2.4.4 The Log Power Spectrogram

After computing the STFT, a spectrogram is typically formed from the power spectrum, as shown in Equation 2.4.

$$P[m,k] = |X[m,k]|^2 \tag{2.4}$$

When working with audio signals, it is common to calculate the logarithm of each vector to create a log power spectrogram as shown in Equation 2.5 [53].

$$S_{log}[m,k] = 20\log_{10} P[m,k] \tag{2.5}$$

This technique is commonly used to aid in visualizing audio, inspired by the way humans perceive sound [54]. In audio recognition tasks, this becomes especially important because spectral energy often resides in the lower frequency ranges [41]. The contrast between the log power spectrogram and the linear power spectrogram is illustrated in Figure 5.

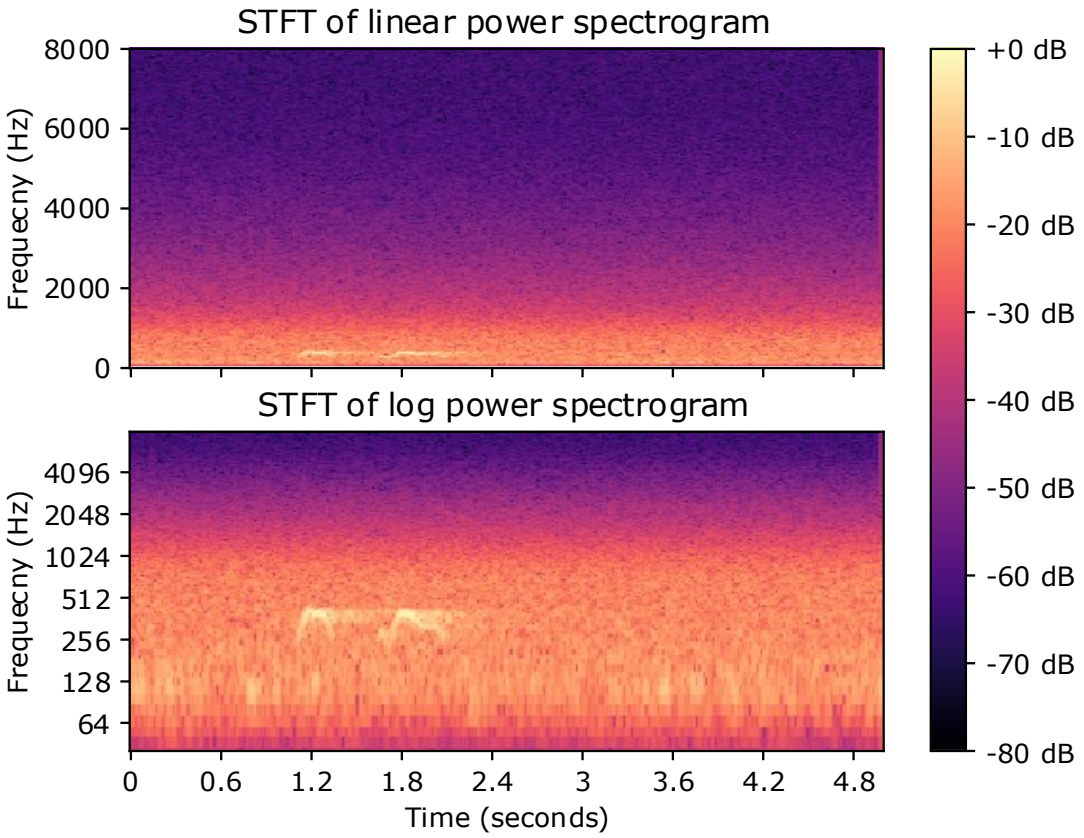

Figure 5: STFT linear spectrogram top and log power bottom

2.4.5 Mel Scale and Mel Spectrograms

The Mel spectrogram is a type of spectrogram in which the frequency axis is adjusted according to the Mel scale, a perceptual scale that approximates the way in which humans perceive pitch [55]. Unlike the traditional linear frequency scale used in standard Short-Time Fourier Transform (STFT) spectrograms, the Mel scale allocates frequency bands non-linearly [56]. This results in higher resolution at lower frequencies and increasingly coarser resolution at higher frequencies. Such a distribution aligns with the human auditory system's heightened sensitivity to differences in lower frequencies, rendering Mel spectrograms particularly effective for analysing natural sounds and biological signals, such as wildlife vocalisations. Figure 6 visualises the Mel scale.

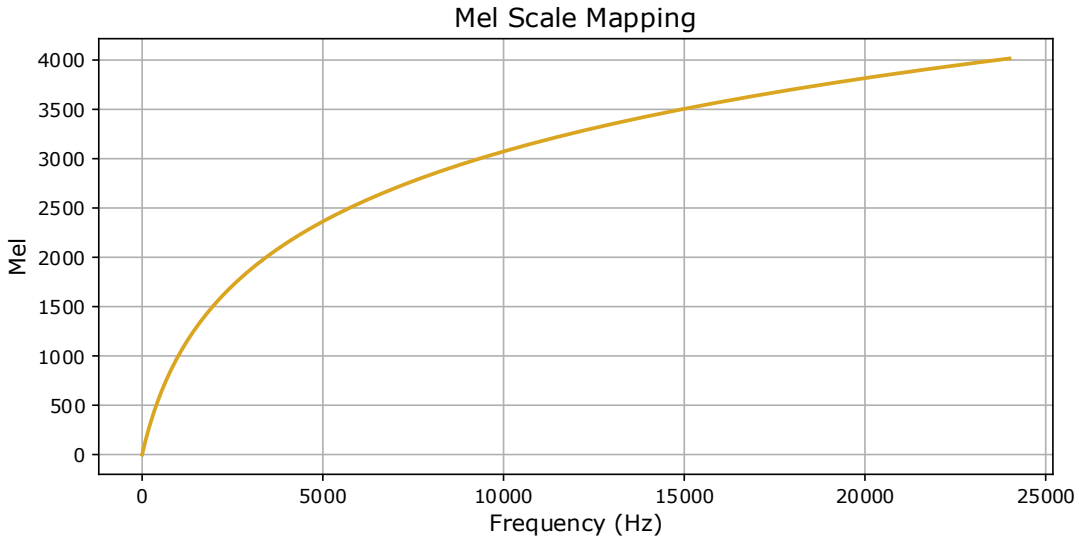

Figure 6: Mel Scale

$$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{2.6}$$

$$f = 700\left(10^{\frac{m}{2595}} - 1\right) \tag{2.7}$$

Equation 2.6 maps frequency from hertz to the Mel scale, and Equation 2.7 converts mels back to hertz. A Mel spectrogram is then created by passing each short-time power spectrum through a bank of overlapping triangular filters with centres evenly spaced on the Mel axis. The Mel filterbank energy is given by Equation 2.8 [57]. For visualisation and dynamic-range control, these band energies are typically log-compressed using Equation 2.9 to produce a decibel-scaled Mel spectrogram [24].

$$M[m, b] = \sum_{k} H_b[k] P[m, k] \tag{2.8}$$

$$\widetilde{M}[m, b] = 20 \log_{10} M[m, b] \tag{2.9}$$

3 Methods

This chapter explains how the thesis applies the theory to the experiments. First, the audio data acquisition and its contents are presented. Next, the processing of the data and spectrogram generation. After this, the deep learning setup and configuration are explained. Finally, tests are described, along with the method of evaluation.

3.1 Tools and Environment

All experiments were implemented in Python 3.10 [58] using *PyTorch* [59] for model training and *Ultralytics* [60] for YOLO implementations. Audio signal processing used *Librosa* [61]. The Figures with maps (Figures 8 and 10) were generated using *Folium* [62]. All training, evaluation, and experiments were conducted on an NVIDIA RTX A6000 GPU (NVIDIA Corporation, Santa Clara, California, USA).

3.2 Data collection and Environment

Acoustic data were collected at 91 fixed recording sites (see Figure 7) within the municipality of Hämeenlinna, Finland. These sites were distributed across two distinct areas. The first area is the Evo hiking region (centred at N 61° 16′, E 25° 09′), which predominantly comprises conserved old-growth forest and is managed by Metsähallitus [63]. Each site was designated based on its location, such as Tuohimetsäntie (Tuohi) and Sudenpesäntie (Su). Most of the data from this region were collected in 2023 and are detailed further in Appendix A. The Lammi Biological Station performed the 2023 data collection in collaboration with the National Land Survey of Finland (NLS) and Huawei.

The second study site encompasses an 8,000-hectare forested area in Harviala, Hämeenlinna, centred at coordinates N 60° 59′, E 24° 43′. This area is owned by UPM and serves commercial purposes. To mitigate habitat loss, several human-made boxes for owls have been installed within the site, supplied by the Finnish Osprey Foundation. All data pertaining to this site were collected in 2025 by researchers at the Lammi Biological Station, including Sukanya Hasan and Nea Hamara. All the locations within this area are named as UPM.

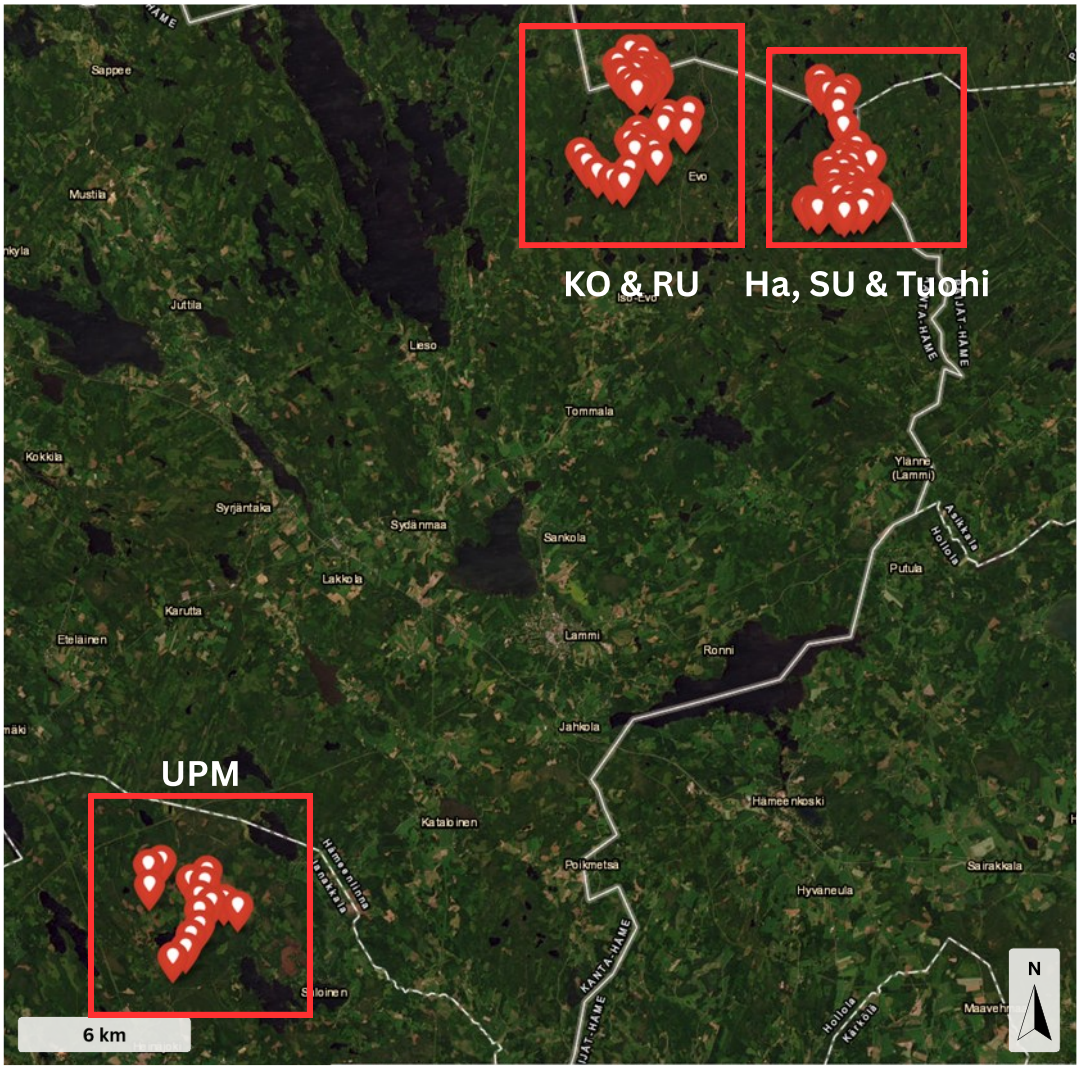

Figure 7: Location of all recording sites. Basemap: @ OpenStreetMap contributors. Contains data from the National Land Survey of Finland. Data available under the Open Database License (ODbL). See https://www.openstreetmap.org/copyright [64].

The Acoustic recordings were captured using the AudioMoth Dev device (Figure 8), a compact and low-power acoustic monitoring platform developed by Open Acoustic Devices (Open Acoustic Devices, Oxford, UK). The device features a Silicon Labs EFM32WG380F256 microcontroller (Silicon Laboratories, Inc., Austin, Texas, USA) and an integrated analog MEMS microphone with a frequency capture range of 10 Hz to 192 kHz, a sensitivity of –38 dBV/Pa, and a signal-to-noise ratio (SNR) of 63 dBA [65]. For this project, the AudioMoth units were configured to capture frequencies up to 48 kHz, sufficient to include all owl vocalisations of interest while managing data volume. Devices were programmed to record 15 to 59-minute audio segments at the beginning of each hour, with rechargeable lithium-ion batteries powering

each unit. Audio data was saved in WAV format to MicroSD cards. All audio files followed a standardised naming, including the site ID, date, and time.

Figure 8: AudioMoth Dev device on the left with an attached battery holder and Li-ion battery on the right.

3.3 Target Owl Species

Figure 9 illustrates the three owl (Strigidae) species examined in this thesis: Ural Owl (Strix uralensis), Boreal Owl (Aegolius funereus), and Eurasian Pygmy Owl (Glaucidium passerinum). The selection of these species was motivated by their shared ecological environments, the conservation concerns of small owl species, their representation in the available datasets, and the presence of interspecific interactions, including predation dynamics [5].

Strix uralensis (Figure 9, A), commonly known as the Ural owl, is distinguished by its notable size and elongated tail, exhibiting a pale grey-brown plumage with dense dark streaking and dark eyes [66]. Strix uralensis is characterised by strong territoriality, particularly around nesting sites, including natural cavities and artificial boxes, demonstrating significant defensive behaviours. Males are primarily responsible for provisioning, while females undertake incubation duties. Notably, fledglings often leave the nest before fully mastering flight. Adult females defend their nesting sites, which is critical for reproductive success. The male Ural owl produces a low, two-part hoot that functions as a territorial call. The diet of Strix uralensis is predominantly composed of voles but diversifies to larger prey during periods of low vole abundance. Furthermore, Strix uralensis poses a considerable threat to

Aegolius funereus, exerting significant predatory pressure, especially during the breeding season.

Glaucidium passerinum (Figure 9, B), recognised as Europe's smallest owl species, features dark grey-brown plumage with fine yellowish spotting above, alongside a pale grey-white underbelly marked by dark streaks and identifiable sulphur-yellow eyes [67]. This species typically exhibits crepuscular activity but also hunts during daylight hours in winter, utilising open perches for rapid, undulating strikes on prey. Nesting typically occurs in small cavities, and Glaucidium passerinum often caches food for later consumption.

Aegolius funereus (Figure 9, C), known as the boreal owl, ranks as the second smallest owl species in Finland, displaying a compact form with a pale, squarish facial disc and prominent yellow eyes adorned with bold white markings [68]. This species employs a concealed roosting strategy during daylight hours and forages predominantly at dusk and throughout the night. Its breeding success is closely tied to the availability of woodpecker (Picidae) cavities or artificial nesting structures, particularly during cycles of abundant vole (Arvicolinae) populations, during which it also engages in food caching. The male boreal owl produces a resonant, flute-like call, frequently heard in late winter and early spring, and the population may exhibit irruptive movements in response to dramatic declines in vole numbers. Due to its smaller stature and reliance on similar prey, Aegolius funereus faces significant predation risk from Strix uralensis, which is one of its principal threats.

Ecologically, all three owl species exhibit a strong reliance on small-mammal populations, particularly voles, resulting in some trophic overlap. However, several factors contribute to niche differentiation among the species, including body size, diel activity patterns, with Glaucidium passerinum showcasing crepuscular behaviour in contrast to the strictly nocturnal tendencies of Aegolius funereus, acoustic vocalisations (which range from the lowest frequencies in Strix uralensis to the highest in Glaucidium passerinum), and varying nesting preferences. Direct interference is most likely to manifest near the nesting sites of Strix uralensis and during periods of low rodent population densities, especially as Aegolius funereus becomes particularly vulnerable to predation during such times.

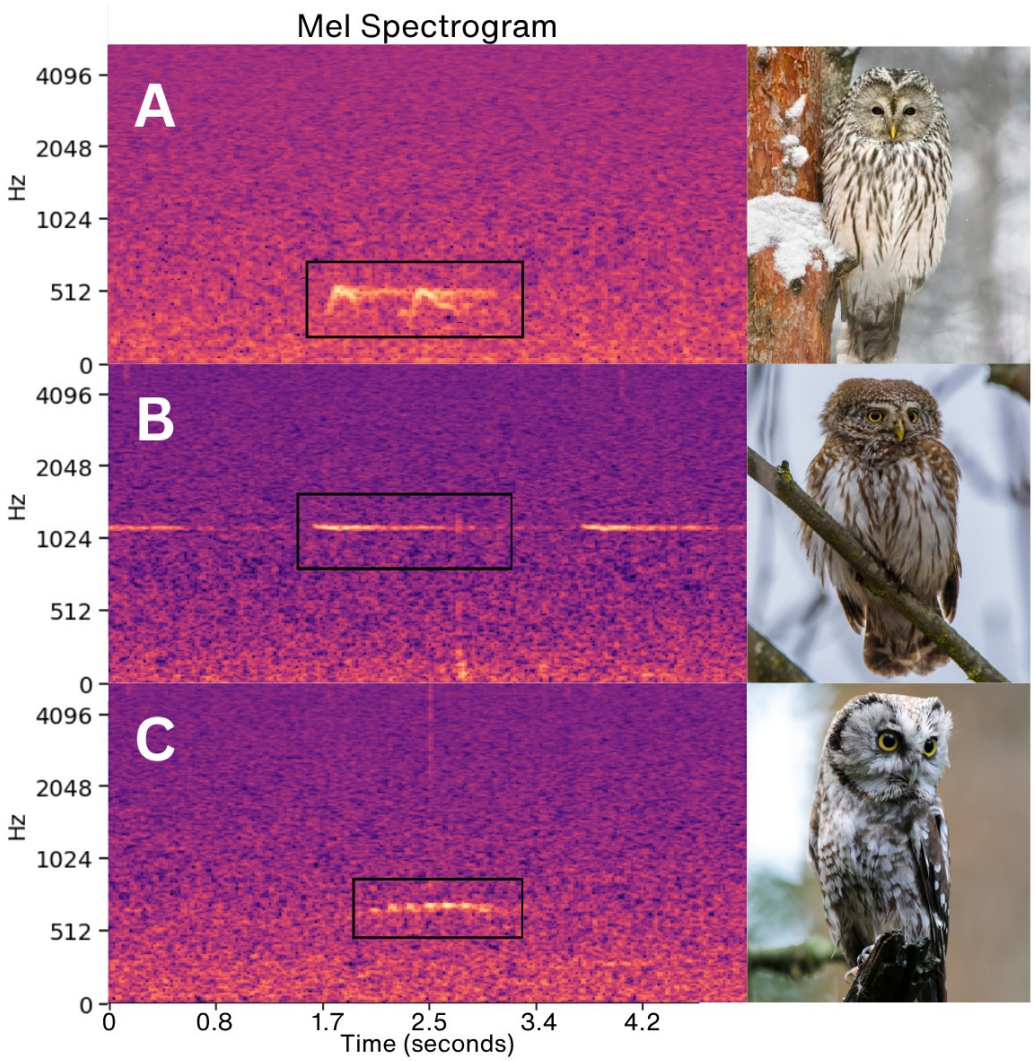

Figure 9: Picture and Mel spectrogram of a call of: (A) Strix uralensis [69], (B) Glaucidium passerinum [70] and Aegolius funereus [71]. Photos [69]-[71] used under Creative Commons licence (CC BY-NC-ND 2.0).

3.4 Signal Pre-processing and Mel Spectrogram Generation

This section elaborates on the methodology used to convert raw audio signals into normalised two-dimensional time-frequency representations, suitable for subsequent object detection tasks.

3.4.1 Segmentation

Each of the original 15-minute recordings was segmented into shorter clips lasting 5 seconds each. This segmentation technique was employed to enhance the temporal resolution of the data, thereby enabling more precise isolation of brief owl vocalisations. Such a methodological approach is commonly adopted in similar studies within the field [34], [72].

3.4.2 Downsampling

The audio segments were downsampled to a sample rate of 8 kHz through the application of an anti-aliasing low-pass filter. In accordance with the Nyquist–Shannon sampling theorem, a sampling rate of 8 kHz sufficiently captures frequency components up to 4 kHz, which encompasses the vocalisation ranges of all owl species examined in this study [50]. Additionally, downsampling reduces computational load and storage demands [73]. Table 1 further illustrates the dominant frequencies characteristic and the Standard Deviation (SD) of each species analysed in this thesis. The dominant frequency was identified by averaging the energy of bird calls across the frequency spectrum.

Species	**Dominant Frequency (Hz) + SD**
Strix Uralensis	512 ± 100
Aegolius funereus	850 ± 113
Glaucidium passerinum	1600 ± 125

Table 1: Dominant frequencies of the species

3.4.3 Generation of Mel Spectrograms

For each five-second audio segment, Mel spectrograms were systematically generated to represent frequency characteristics over time. These spectrograms were computed utilising the *Librosa* audio processing library, employing 256 Mel bands to ensure sufficient resolution in the frequency domain for detecting subtle variations in owl calls. An FFT window size of 2048 samples was chosen to balance frequency resolution with computational efficiency. A hop length of 512 samples was employed, representing the number of samples between successive windows; given the window size, this results in approximately 75% overlap. This overlap enhances temporal resolution while maintaining manageable computational load. The Mel spectrograms were subsequently converted to a decibel (dB) scale using *Librosa's* power_to_db function, allowing for more accurate visual and analytical comparisons of spectral features. All resultant values are detailed in Table 2.

Parameter	Value
Length of the recordings	5 s
Sampling Rate	8000 Hz
Window length	2048 samples
Hop Length	512 samples
Scale	Log-power (dB)

Table 2: Mel spectrogram parameters

3.5 Annotation and Dataset Construction

This section describes how the previously generated Mel spectrograms were annotated. It then presents the contents of two datasets, one for training a model and one for deploying it.

3.5.1 Label set and annotation guidelines

Owl vocalisations were manually annotated within the Mel spectrograms. For this, the bounding-box method was used to identify regions of the spectrogram where vocal activity was present. Overlapping calls were annotated as separate instances. The label set comprised Strix uralensis, Aegolius funereus, Glaucidium passerinum, and noise. Noise didn't contain any of the three species mentioned above and was used to improve the results against false positives. The correctness of some labels was confirmed with experts in the field. Labels were generated using Label Studio (Heartex, Inc., San Francisco, California, USA) and exported in COCO format [74] for Faster R-CNN and in TXT format for YOLO.

3.5.2 Dataset Content

Two separate datasets were constructed to fulfil the objective of this thesis: to develop and compare object-detection-based identification methods for owl species and to create an end-to-end pipeline based on them. The first dataset comprises data collected from various sites, with no restrictions on collection date or geographic location. The primary criterion for inclusion was the availability of clearly labelled identifiable data. A second criterion was that Dataset 1 did not contain any recordings overlapping with those in Dataset 2. In total, 28 sites from 2023 and 15 sites from 2025 were utilised (see Appendix A). The 2023 sites were selected based on owl presence, as indicated in Arbimon [75], where all 2023 data were uploaded. Of the 2025 sites, all were used, excluding those in dataset 2.

In more detail, the first dataset consists of manually annotated Mel spectrogram images, each labelled with one or more owl species or as background noise. Each instance corresponds to a unique bounding box annotation within a 5-second spectrogram. A 5-second Mel spectrogram may depict multiple instances corresponding to distinct vocalisations if more than one vocal event occurs within the interval. In the absence of vocal activity, the spectrogram displays noise, with no indication of any of the three owl species. Consequently, no instances of owl vocalisations are present during such noise segments. The dataset was partitioned employing a stratified split, allocating 80% to the training set, 10% to the validation set, and 10% to the test set. This approach ensures that each subset maintains a proportional representation of each class. The number of images, instances, and the remaining images after the split can be seen in Tables 3 and 4.

Label	**Images**	**Instances**
Strix Uralensis	1194	1763
Aegolius funereus	1025	1431
Glaucidium passerinum	1833	4082
Noise	1235	0
Total	5287	7276

Table 3: Dataset 1, segments and Instances

Dataset	**Images**
Train set	3458
Validation set	568
Test set	568

Table 4: Dataset 1 size and distribution

The second dataset contains only raw audio files for the second experiment. It includes 12 different sites over 15 days, from February 18, 2025, to March 4, 2025. Each site has recordings for 59 minutes every hour, totalling 354 hours per site and 4248 hours total. All UPM sites are shown in Figure 10, and the two Tuohi sites are visible in Figure 7.

Figure 10: 10 UPM area sites of dataset 2. Basemap: @ OpenStreetMap contributors. Contains data from the National Land Survey of Finland. Data available under the Open Database License (ODbL). See https://www.openstreetmap.org/copyright [64].

Dataset 2	Value
Total sites	12
Total length of recordings (h)	4248
Date	18.02.2025-04.03.2025

Table 5: Dataset 2

3.6 CNN Architecture and Training Information

This section explains how the models were configured and evaluated. This includes the number of epochs, batch sizes, optimizers, and learning rates.

3.6.1 Faster R-CNN ResNet-50 and YOLOv11

This thesis aimed to develop a transfer learning-based CNN for object detection, with a specific focus on identifying owl calls from spectrogram data. Two architectures were chosen: Faster R-CNN [18] with a ResNet-50 [12] backbone and YOLOv11-l [13]. The choice of Faster R-CNN is supported by extensive prior validation of the ResNet-50 architecture, which has demonstrated a reliable balance between accuracy and model capacity when applied to spectrogram inputs [33], [76]. Faster R-CNN enables the use of ResNet-50 as the backbone while producing bounding box outputs. In contrast, YOLO represents a one-stage detection framework that is widely adopted in real-time applications, renowned for its high recall and low latency. Although YOLO-based architectures [13] are less commonly employed for this particular application, existing prior work indicates their efficacy [41].

To ensure a fair comparison, the YOLOv11-l variant was chosen because it has a similar number of parameters to the Faster R-CNN architecture using ResNet-50. This helps control for model complexity and isolates the effect of architectural differences. Another motivation for choosing these two architectures was to compare one-stage and two-stage architectures, especially their effectiveness in localising events within the time-frequency domain and their inference speeds, which is especially important given the large data volume and the need for quick processing. Both detector families also come with mature, well-supported, pre-trained weights that have demonstrated strong performance across various applications, providing reliable benchmarks for our evaluation.

In summary, both architectures are engineered to perform similar tasks, yet they exhibit significant differences. This distinction enables a comprehensive evaluation of object detection-based methodologies for identifying avian species through vocalisations.

3.6.2 Hyperparameters and configurations

In this section, the hyperparameters and configurations used for two models, YOLOv11 and Faster R-CNN with a ResNet-50 backbone, are detailed. Both models were configured to process images of the same dimensions. The datasets were formatted appropriately for each model, with YOLOv11 utilising labels in the PyTorch TXT format, while Faster R-CNN employed the COCO

dataset format. To ensure a fair comparison, similar default weights were maintained across both models. The number of training epochs was adjusted individually for each architecture to achieve convergence. The YOLOv11-l model was selected, containing fewer parameters (23.3 million) compared to the Faster R-CNN ResNet-50 model, which has 41.8 million parameters. The AdamW optimizer was used for both models, given its proven efficacy in previous research [77]. The learning rate for Faster R-CNN was set lower because it became unstable with a higher learning rate. All specific parameter settings are detailed in Table 6.

Parameter	Faster R-CNN	YOLOv11-l
Input size	640 x 640	640 x 640
Optimizer	AdamW	AdamW
Epoch	90	150
Batch size	16	16
Learning rate	0.0001	0.001
Weight decay	0.0005	0.0005
Parameters	41.8 M	25.3 M

Table 6: Model parameters for Faster R-CNN, YOLOv11-l. Hyperparameters that differ between models are highlighted grey

3.7 Experimental Design and Evaluation

This section introduces the design and evaluation of two key experiments in the thesis. The first experiment aims to compare the effectiveness of two advanced deep learning models for detecting owl vocalisations from Mel spectrograms, focusing on model performance metrics. The second experiment applies the optimal model to a larger, unlabelled dataset across multiple sites to assess its scalability and practical utility in real-world bioacoustics monitoring. Detailed methodologies and metrics are discussed in the following sections.

3.7.1 Experiment 1: Comparative evaluation of detectors

In this study, a comparative analysis was conducted to evaluate the performance of two prominent object detection models: Faster R-CNN with a ResNet-50 backbone and YOLOv11-l. Both models were trained using the specified hyperparameters until convergence was achieved, ensuring optimal performance for each architecture. Following the training phase, a series of evaluation protocols were implemented to assess and compare the models' accuracy and classification speed, thereby providing a comprehensive understanding of their capabilities within the context of this research.

The model's performance was assessed using five standard classification metrics: Accuracy, Precision, Recall, F1 Score, and mAP. mAP was evaluated using two metrics, mAP50 and mAP50:95. These metrics provide insights into the model's overall accuracy and its ability to minimise false positives and false negatives. Additionally, the inference time was measured to determine the model's suitability for practical applications.

Accuracy quantifies the overall performance of the model by assessing the proportion of correctly classified instances relative to the total number of instances. It is mathematically defined as:

$$Accuracy = \frac{Correct\ Predictions}{Total\ Predictions} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.1}$$

True Positives (TP) refer to instances correctly identified as belonging to a specific class, whereas False Positives (FP) denote instances incorrectly predicted as members of that class. True Negatives (TN) represent correctly identified non-members of the class, while False Negatives (FN) indicate true class members that the model failed to recognize.

Precision is a metric that quantifies the proportion of true positive predictions relative to all positive predictions made by the model. It serves as an indicator of the model's capacity to minimize false positives, thereby ensuring that predictions of the positive class are indeed correct. Formally, it is defined as:

$$Precision = \frac{TP}{TP + FP} \tag{3.2}$$

Recall quantifies the model's capacity to identify all relevant objects within a given dataset. Mathematically, it is expressed as the ratio of true positive detections to the total number of actual objects present. A high recall value signifies that the model effectively detects the majority of objects, although it does not provide information regarding the prevalence of false positives. Thus, recall is defined as:

$$Recall = \frac{TP}{TP + FN} \tag{3.3}$$

The F1 score is a statistical measure that represents the harmonic mean of precision and recall. It serves as an indicator of the balance between these two metrics. Formally, it is defined as:

$$F_1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{3.4}$$

Mean Average Precision (mAP) is a comprehensive metric for evaluating the performance of object detection models. It summarises the precision-recall curve by calculating average precision (AP) across all recall levels for each object class, then averaging these values across all classes. This metric provides an overall measure of the detector's ability to accurately identify and localise objects in images, balancing precision and recall.

$$mAP = \frac{1}{N} * \sum_{i=1}^{N} (AP_i) \tag{3.5}$$

In the context of mAP, there are specific variations such as mAP50 and mAP50:95. mAP50 measures the average precision at an Intersection over Union (IoU) threshold of 0.5. This means that a predicted bounding box is considered correct if its IoU with the ground truth box is at least 0.5. This metric provides a baseline evaluation of the model's detection ability at a lenient threshold.

mAP50:95 provides a more comprehensive evaluation by calculating average precision across multiple IoU thresholds (0.5-0.95 in increments of 0.05). This metric emphasises the model's performance across a range of stricter detection criteria. As a result, mAP50:95 typically provides a more rigorous assessment of the model's ability to localise objects accurately [78].

Finally, the classification time is assessed through two distinct metrics to ensure reproducibility. The first metric, per-image inference latency, measures the time taken for a single forward pass plus post-processing, reported in milliseconds per image. The second metric, end-to-end pipeline time, captures the total wall-clock time required to process an entire 59-minute recording, including preprocessing, detector inference, and all other processing steps, reported in seconds per file.

3.7.2 Experiment 2: Deployment of the mode on raw data

This experiment utilised the best-performing detector identified in Experiment 1 to analyse raw recordings collected over the period from February 18, 2025, to March 4, 2025. Only twelve sites were incorporated into the study, as these were the only locations that met the criteria for synchronous recording windows throughout the entire study period. Sites with inconsistent recording times or missing data on certain days were excluded to ensure comparability across the study's temporal span.

The deployment consisted of non-overlapping five-second clips extracted from the original recordings. These clips were processed using the same pre-processing and Mel-spectrogram generation pipeline described in Sections 3.2 and 3.3, ensuring consistency across the dataset. Timestamps were parsed directly from audio record filenames and normalised to Helsinki local time (UTC+2).

Inference was performed using the architecture and checkpoint selected in Experiment 1. The evaluation parameters are detailed in Table 7. Confidence refers to the model's certainty about the presence of an object within a predicted bounding box [13]. It is represented as a score between 0 and 1, with higher scores indicating greater confidence. Confidence is calculated in the final layer using softmax [79]. Each clip was processed once, and for each detection, the timestamp, date, species label, confidence score, and site identifier were recorded. No additional post-processing steps were applied.

The detection results were summarised to facilitate downstream analysis. Three types of outputs were generated for the period from February 18 to March 4, 2025. These included: daily detections by species, hourly detections by species, and total detections by site, along with species composition.

Parameters	Value
Image Size	640 x 640
Confidence	0.7
IoU	0.45

Table 7: Evaluation Parameters

4 Results

4.1 Experiment 1: Model Performances

This section reports the results of Experiment 1 in the thesis. It covers the performance of Faster R-CNN and YOLOv11-l. The results include metrics such as mean Average Precision (mAP), accuracy, precision, recall, F1-score, and inference speed.

4.1.1 YOLOv11-l

YOLOv11-l was trained for 150 epochs, which took 220 minutes. Figure 11 shows the loss, mAP50 and mAP50:95 for the model during training. The mean average precisions for each class are reported in Table 8. On the test set, the model achieved an overall mAP50 of 99.0% and mAP50:95 of 75.7%.

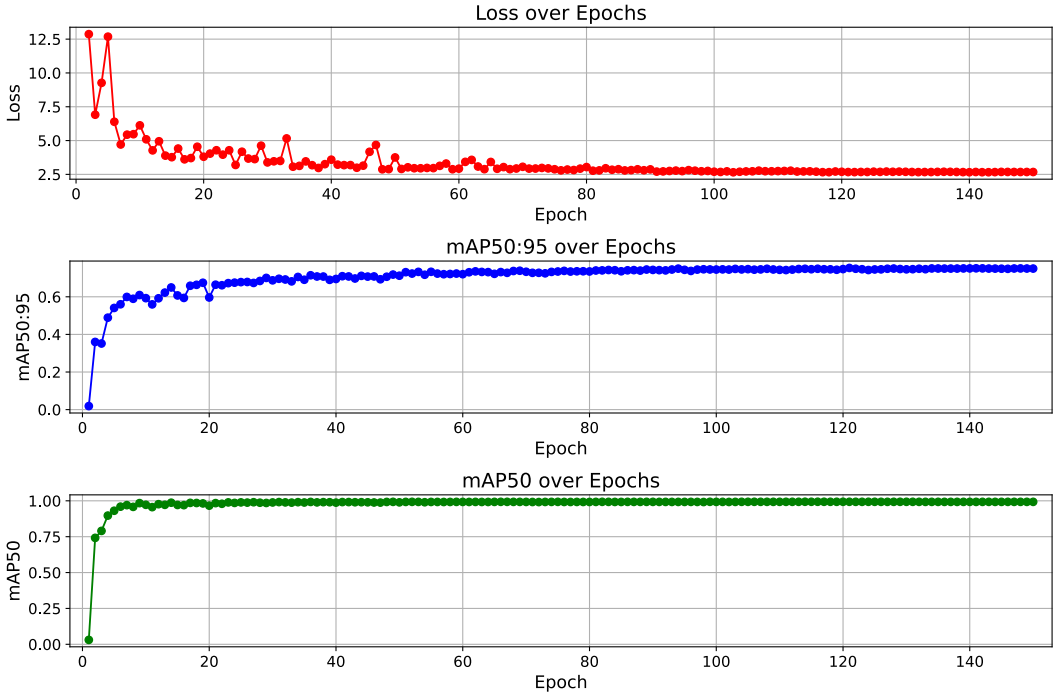

Figure 11: Training loss, mAP50 and mAP50:95 for YOLOv11-l over 150 epoch

Species	mAP50	mAP50:95
All	99.0 %	75.7 %
Aegolius funereus	99.0 %	76.1 %
Glaucidium passerinum	98.8 %	78.6 %
Strix uralensis	99.2 %	72.4 %

Table 8: mAP50 and mAP50:95 for each species on the test set

When evaluated on the test dataset, the model attained an accuracy of 93.8% and an F1 score of 96.8%, as depicted in Table 9. The average inference latency was measured at 7.2 milliseconds per image, corresponding to approximately 139 frames per second. The confusion matrix presented in Figure 12 delineates the comparison between predicted and actual class labels.

Evaluation Parameter	Score
Accuracy	93.8 %
Precision	94.3 %
Recall	99.5 %
F1 Score	96.8 %
Inference Speed	7.2 ms
Processing speed	52.8 s per hour of data

Table 9: Evaluation results with YOLOv11-l

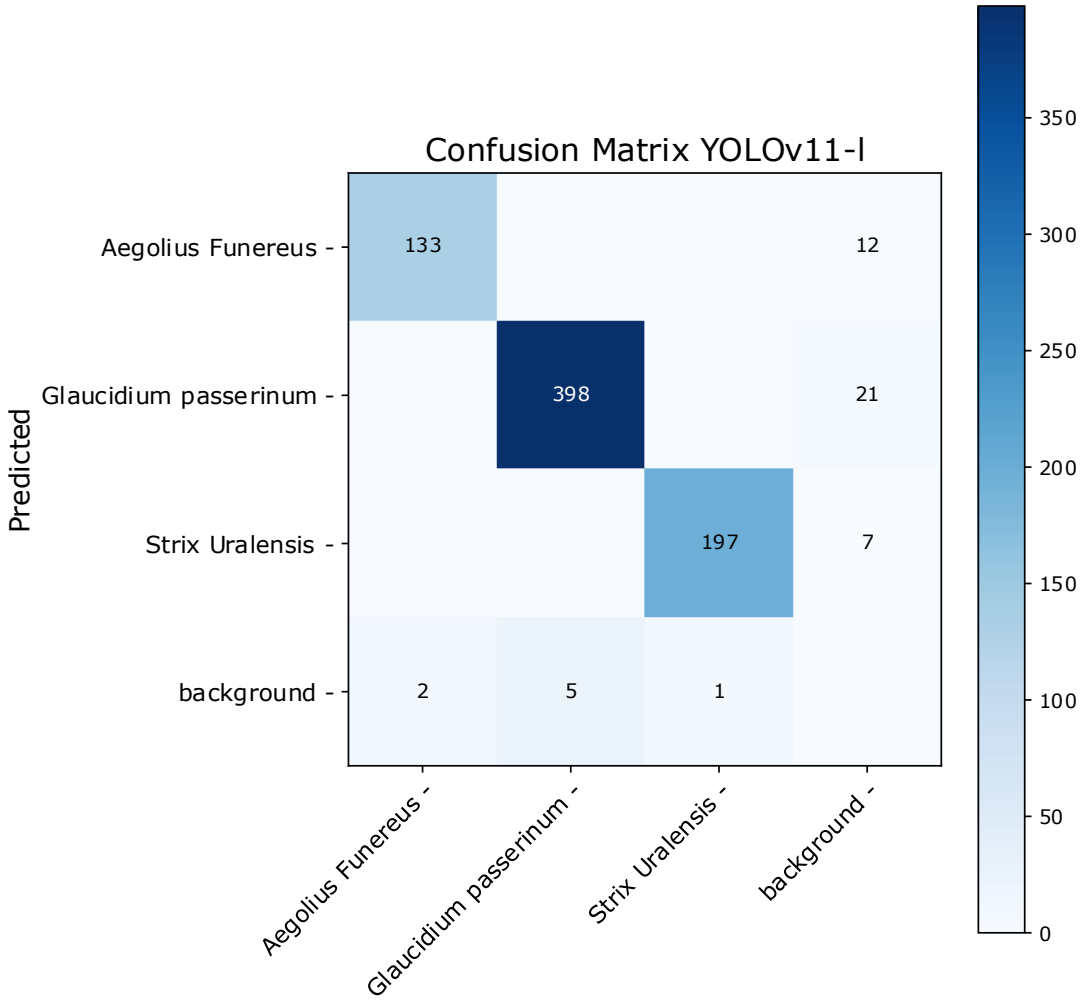

Figure 12: Confusion matrix with YOLOv11-l on the test set

4.2 Faster R-CNN Resnet-50

Faster R-CNN with Resnet-50 backbone was trained for 90 epochs, which took 350 minutes. Figure 13 shows the loss, mAP50 and mAP50:95 for the model during training. The average precision for each class is reported in Table 10. The final model achieved a mAP50 of 97.2% and a mAP50:95 of 66.9% on the test set.

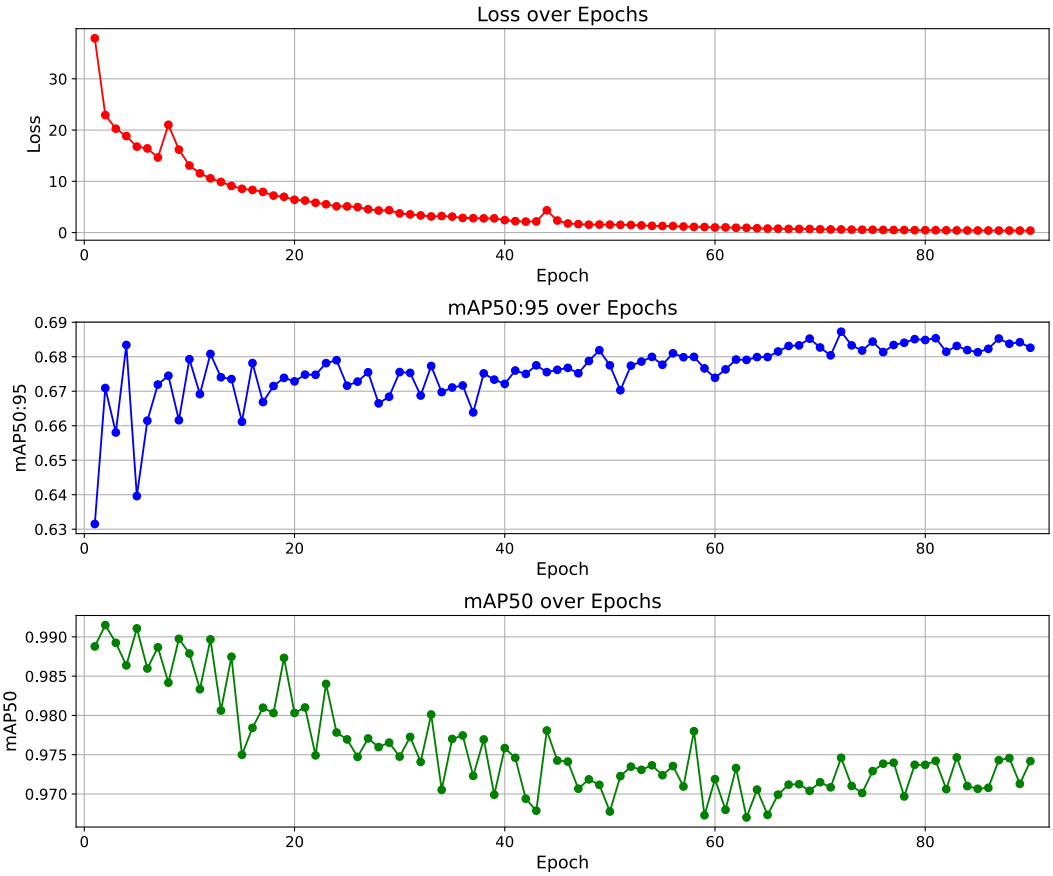

Figure 13: Training loss, mAP50 and mAP50:95 for Faster R-CNN over 90 epoch

Species	**mAP50**	**mAP50:95**
All	97.2 %	66.9 %
Aegolius funereus	96.3 %	61.7 %
Glaucidium passerinum	97.8 %	71.5 %
Strix uralensis	97.7 %	67.6 %

Table 10: mAP50 and mAP50:95 for each species on the test set

When evaluated on the test dataset, the model attained an accuracy of 92.8% and an F1 score of 96.3%, as depicted in Table 11. The average inference latency was measured at 29.4 milliseconds per image, corresponding to approximately 34 frames per second. The confusion matrix presented in Figure 14 delineates the comparison between predicted and actual class labels.

Evaluation Parameter	Score
Accuracy	92.8 %
Precision	97.5 %
Recall	95.1 %
F1 Score	96.3 %
Inference Speed	29.4 ms
Processing Speed	81.5 s per hour of data

Table 11: Evaluation results with Faster R-CNN

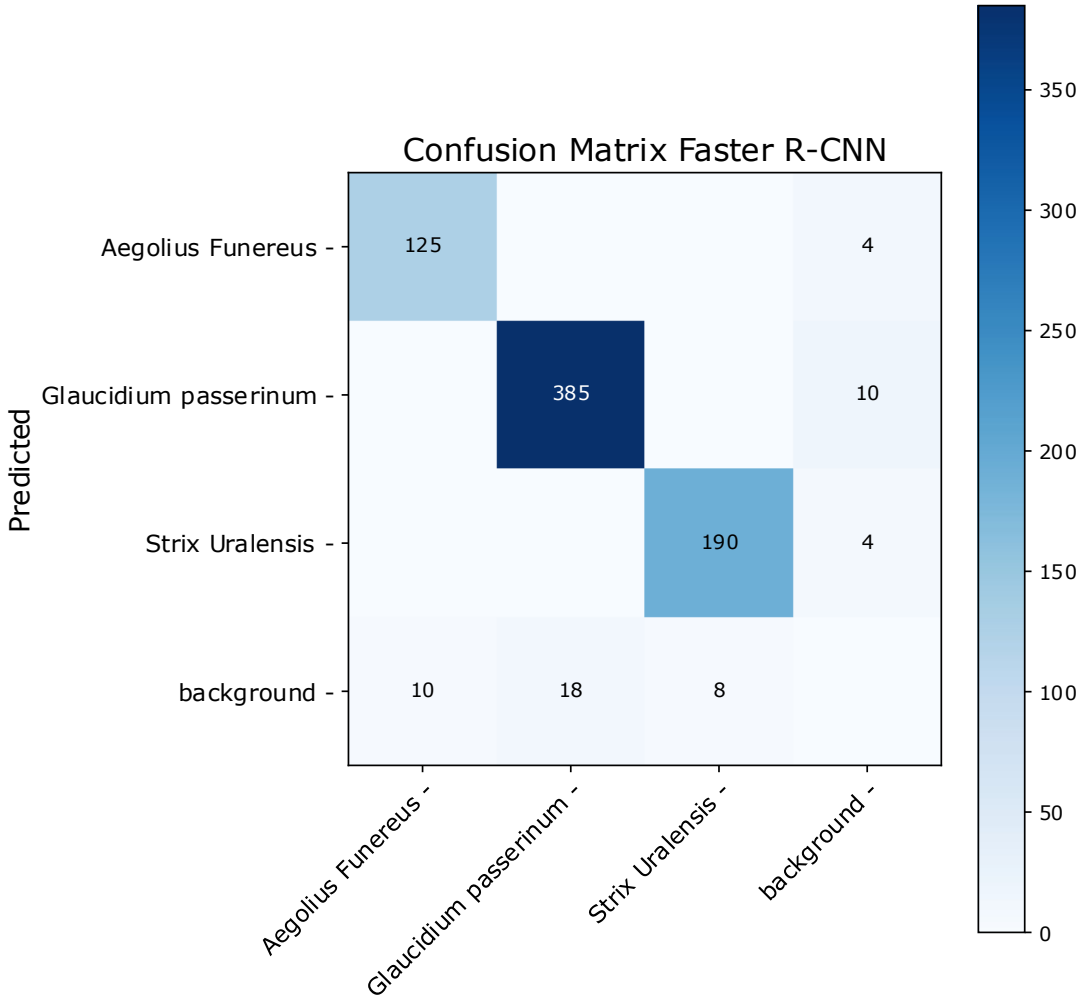

Figure 14: Confusion matrix with Faster R-CNN on the test set

4.3 Experiment 2: Application on Large Dataset

This section describes the detection of three owl species utilising the YOLOv11 pipeline. The process required a total of 61.25 hours and yielded 12,096 detections, as summarised in Table 12. Results are presented through various analytical perspectives, including site, hour, and date.

Parameter	Value
Total classifying time (hours)	61.25
Glaucidium passerinum detections	8111
Aegolius funereus detections	1940
Strix uralensis detections	2045
Total detections	12096

Table 12: Detection results

Figure 15 illustrates the detections recorded across various sites, organised by species. Glaucidium passerinum was consistently detected at all sites, with the highest detection rates observed at Upm 9, Upm 10, and Upm 21. Strix uralensis detections were predominantly confined to Tuohi 1, with fewer occurrences at Tuohi 3. Aegolius funereus showed fewer total detections overall.

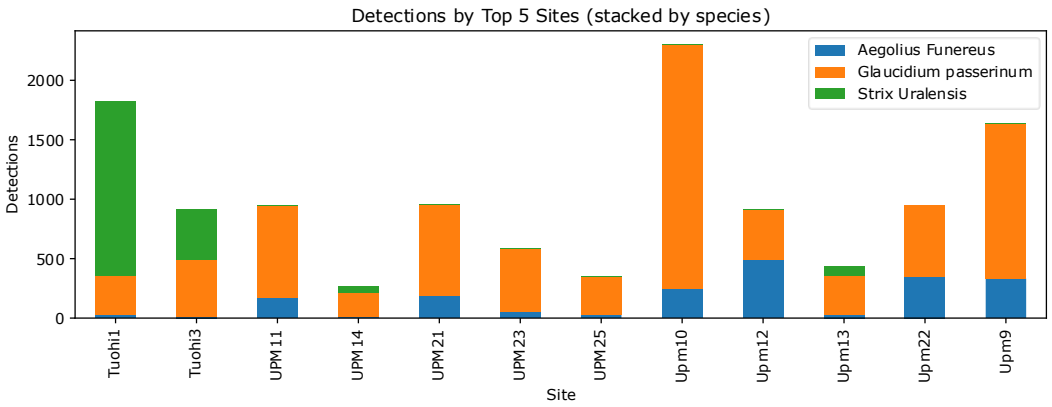

Figure 15: Detections by sites (stacked by species) with the YOLO11 pipeline

Hourly detection data depicted in Figure 16 illustrates the activity patterns of three owl species over the period from February 18 to March 4, 2025. Glaucidium passerinum primarily exhibited activity during the early morning hours between 05:00 and 07:00, as well as in the evening from 17:00 to 19:00. Following sunrise, detections declined during daylight hours, with significantly fewer sightings recorded between 20:00 and 04:00. In contrast,

Aegolius funereus demonstrated a more evenly distributed detection pattern from evening through to morning, with a notable decrease in activity between 17:00 and 09:00. Strix uralensis was predominantly detected from 18:00 to 06:00, with minimal detections occurring during daylight hours.

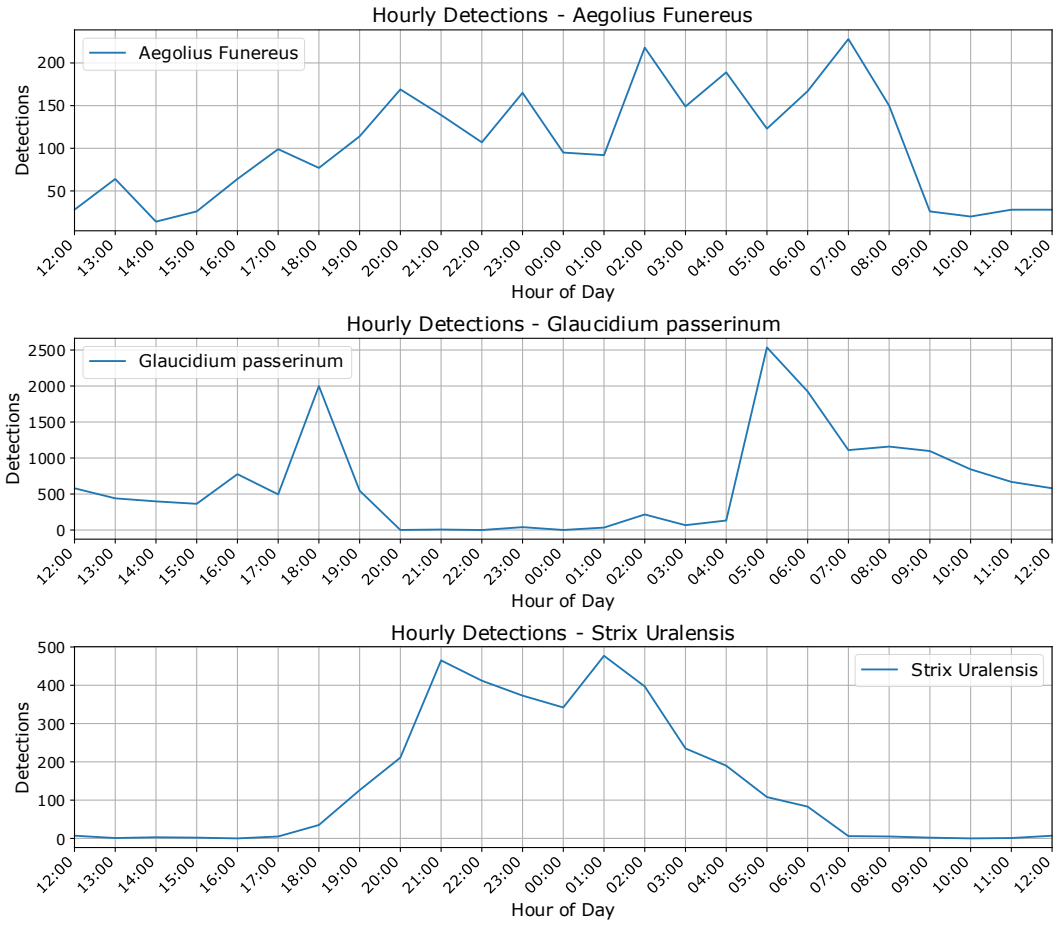

Figure 16: Hourly detections for Glaucidium passerinum, Aegolius funereus and Strix uralensis between 18 February 2025 and 04 March 2025.

As illustrated in Figure 17, the daily detection counts exhibited significant variation throughout the study period from 18 February 2025 to 04 March 2025. Notably, Glaucidium passerinum demonstrated the most pronounced fluctuation, with a peak of approximately 3,200 detections on February 28, followed by a decline to roughly 1,100 on March 1 and subsequent lower counts. Aegolius funereus maintained a generally lower detection rate, with sporadic detections on most days and a maximum of approximately 450 on March 3. Strix uralensis showed two distinct peaks, around 400 detections on February 26 and approximately 700 on March 4, interspersed with several days registering near zero detections.

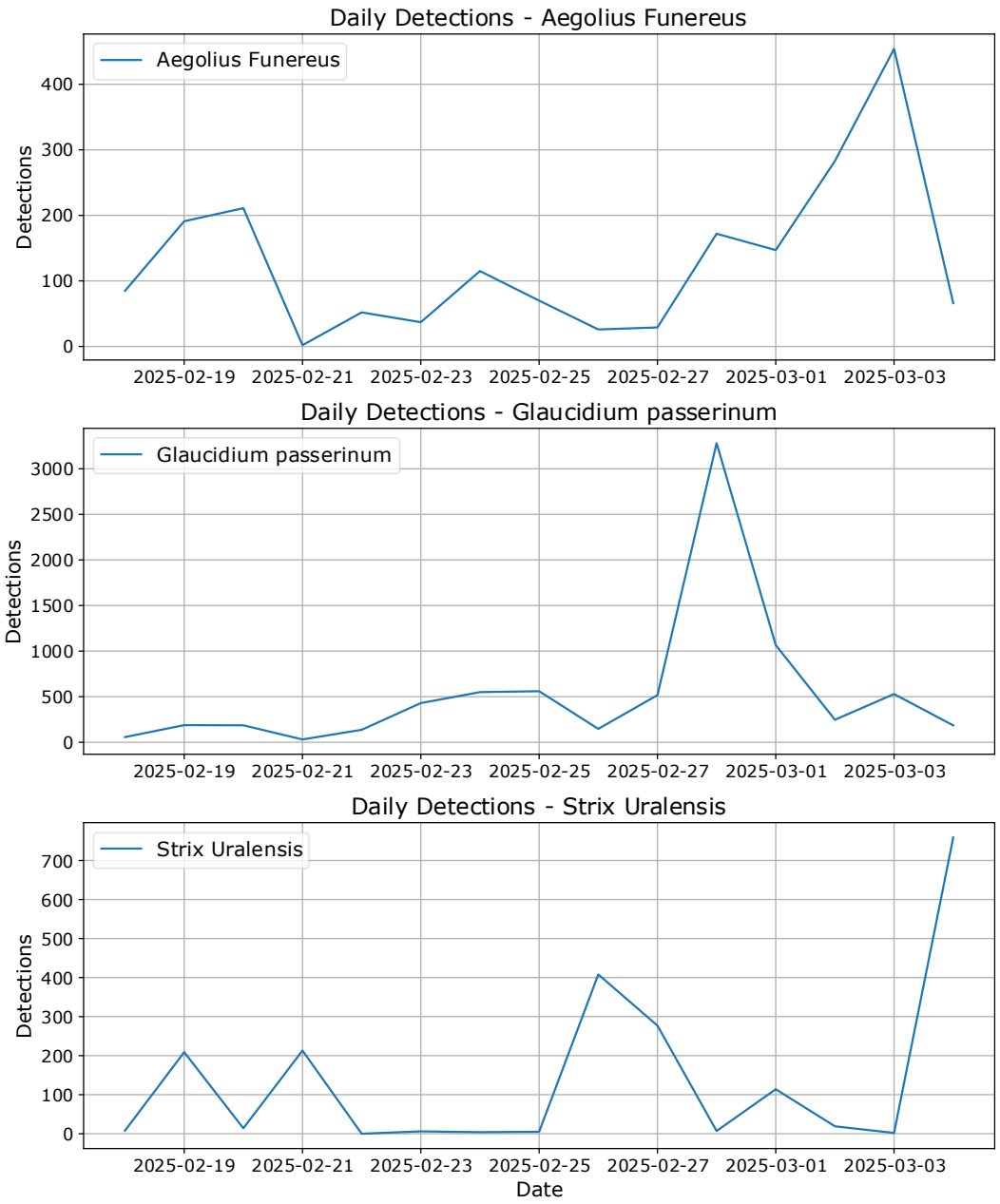

Figure 17: Detections by date for Glaucidium passerinum, Aegolius funereus and Strix uralensis

5 Discussion

5.1 Model Analysis and Performance Comparison

The performance of the YOLOv11-l model during training is illustrated in Figure 11. Throughout the training process, the loss function showed a steady decline, indicating effective learning. Both metrics, mAP50 and mAP50:95, demonstrated consistent improvement, with convergence achieved around epoch 50 when the optimal weights were saved. Evaluation on the test set revealed an impressive mAP50 of 99.0% and an mAP50:95 of 75.7%. Notably, Glaucidium passerinum excelled with the highest mAP50:95 score, likely due to its simpler call pattern, as shown in Figure 9, while Strix uralensis had a lower score of 72.4%, reflecting its more complex call signals.

In addition to these metrics, Table 9 outlines further evaluation parameters, including accuracy, precision, recall, F1 score, inference speed, and processing speed. YOLOv11 achieved an accuracy of 93.8%, with a recall of 99.5%, indicating effective identification of true positives. However, the precision of 94.3% suggests a presence of false positives. The model's inference speed was noted at 7.2 milliseconds per image. The entire processing speed for one hour of raw data was recorded at 52.8 seconds. The confusion matrix in Figure 12 provides a visual comparison between predicted and actual class labels.

Conversely, Figure 13 depicts the training performance of the Faster R-CNN model with a ResNet-50 backbone. The loss curve stabilised around the 60th epoch, indicating convergence, while the mAP50:95 continued to improve. However, the mAP50 exhibited a slight decline, likely due to enhanced localisation accuracy at higher IoU thresholds, which resulted in fewer loosely detected bounding boxes. Evaluation results showed an overall mAP50 of 97.2% and an mAP50:95 of 66.9%. In Table 10, Glaucidium passerinum again topped with an mAP50:95 score of 71.5%, while Strix uralensis and Aegolius funereus had lower scores of 67.6% and 61.7%, respectively.

Table 11 provides additional evaluation metrics for the Faster R-CNN model. A comparative analysis of the detection metrics and inference latency between the two models is presented in Table 13. The YOLOv11 model outperformed Faster R-CNN in mAP50:95, achieving 75.7%, compared to Faster R-CNN's 66.9%, resulting in an 8.8 percentage point difference. Additionally, YOLOv11 demonstrated a significantly faster inference time of 7.2 milliseconds per image, compared to 29.4 milliseconds for Faster R-CNN, resulting in a 22.2-millisecond difference. When considering end-to-end throughput,

as detailed in Table 14, a substantial 35.2% difference in overall processing speed was noted.

Regarding overall accuracy, YOLOv11-l achieved 93.8%, slightly exceeding Faster R-CNN's 92.8%. While Faster R-CNN achieved a higher precision of 97.5% compared to YOLOv11's 94.3%, the recall was notably higher in YOLOv11 at 99.5% versus 94.3%. The overall F1 scores of both models were comparable, with YOLOv11 attaining 96.8% and Faster R-CNN reaching 96.3%. These findings align with previous research comparing YOLO and Faster R-CNN architectures [45]. Considering all these factors, YOLOv11-l was selected for Experiment 2.

Model	mAP@0.5-0.95	Inference(ms)	Accuracy	Precision	Recall	F1-Score
YOLO11	75.7 %	7.2	93.8 %	94.3 %	99.5 %	96.8 %
Faster R-CNN	66.9 %	29.4	92.8 %	97.5 %	95.1 %	96.3 %

Table 13: Performance analysis of YOLO11 and Faster R-CNN

Model	Processing speed (s per 1 hour of audio)
YOLO11	52.8 s
Faster R-CNN	81.5 s

Table 14: Processing speed average for ten 1-hour files in the end-to-end pipeline

In both architectures, interspecific confusions do not occur (see Figures 12 and 14), indicating that the learned classes for the three owl species are well separated within the time-frequency domain. False detections are solely attributable to false positives and negatives caused by factors such as noise, weather conditions, and the presence of other species outside the scope of this study.

To contextualise the findings presented in this thesis within the existing landscape of avian sound recognition methodologies, Table 15 delineates the key attributes and F1 scores of three similar approaches: BirdNET [16], the Faster R-CNN system proposed by Shrestha et al. [14], and the YOLO-BEM detector [41], alongside the introduced YOLOv11-l model. A discernible trend is evident: as the number of target species increases and label quality diminishes, there is a corresponding decline in F1 scores. Specifically, BirdNET, which was trained on a dataset covering 984 species with clip-level labelling and a

3-second input window, achieves an F1 score of 50.0%. In contrast, the Faster R-CNN model, which targets 50 species using a 2-second bounding box label scheme, achieves a score of 85.0%. The YOLO-BEM detector, capable of recognising over ten species with bounding-box supervision, improves performance to 95.0%. Notably, the presented YOLOv11-l model achieves the highest F1 score of 96.8%, utilising approximately 25 million parameters, similar to YOLO-BEM but significantly fewer than Faster R-CNN, which has around 42 million. However, it is essential to remember that this high performance was achieved on a more straightforward task involving just three species, supported by strongly labelled 5-second spectrograms.

Model	Spectrogram length	Number of species	Labelling	Parameters	F1 score
BirdNET [16]	3 s	984	Clip-level	~27 M	50.0 %
Faster R-CNN [14]	2 s	50	Bounding box	~42 M	85.0 %
YOLO-BEM [41]	-	10+	Bounding box	~26 M	95.0 %
presented YOLOv11-l	5 s	3	Bounding box	~25 M	96.8 %

Table 15: Comparison of related solutions

5.2 Analysing results from Large Dataset

Building upon the findings from Experiment 1, the highest-performing detector, YOLOv11-l, was utilised to analyse Dataset 2 (see Table 5). A standardised confidence threshold of 0.7 and an Intersection over Union (IoU) threshold of 0.45, as outlined in Table 7, were consistently applied to all analyses to ensure the comparability of results. These parameters were selected to exclude low-probability detections, with most reliable detections exhibiting confidence scores exceeding 0.8. Detection outcomes were examined by site, hour, and date, allowing for an investigation into spatial and temporal detection patterns. This methodological framework facilitated a comprehensive characterisation of detection structures while maintaining uniform decision criteria throughout the study.

In Experiment 2, 4248 hours of audio data were processed in approximately 61 hours, demonstrating the high feasibility of this approach for large-scale deployments. Furthermore, the methodology can be scaled effectively by upgrading the processing equipment. Among the species recorded, Glaucidium passerinum emerged as the most frequently detected, with a significant presence across all monitored sites, particularly at UPM10, UPM21, and UPM22. Strix uralensis is nearly absent in the UPM region, likely due to the predominance of commercial forest management practices that reduce the

availability of natural nesting sites for this species. In contrast, Strix uralensis was detected in Tuohi 1 and Tuohi 3, located in the EVO old-growth forest. Aegolius funereus was the least represented species within this dataset, with minimal detection instances mainly concentrated in the UPM area. It is also worth noting that when Strix uralensis is present (as shown in Figure 15), detections of Aegolius funereus are almost non-existent.

The hourly detection data presented in Figure 16 delineate clear diel activity patterns that generally correspond with the established behavioural ecology of the species studied. Glaucidium passerinum predominantly exhibited activity during the crepuscular hours, with peak detections occurring between 05:00 and 07:00, and a secondary activity period from 17:00 to 19:00. Post-sunrise, detections markedly declined during daylight hours, with negligible sightings recorded between 20:00 and 04:00, consistent with its classification as a crepuscular species. Conversely, Aegolius funereus displayed a primarily nocturnal activity pattern, with most detections concentrated from 17:00 to 09:00, and a significant reduction in detections after 09:00. Strix uralensis, also characterised as nocturnal, exhibited an even more nocturnal activity pattern compared to Aegolius funereus, with detections primarily between 18:00 and 06:00, particularly focusing around 21:00 and 02:00. During daylight hours, detections were minimal.

Figure 17 illustrates considerable daily variability throughout the study period from 18 February to 4 March 2025. The species Glaucidium passerinum demonstrates the most pronounced fluctuations, reaching a peak of approximately 3,200 detections on February 28, followed by a decline to around 1,100 detections on March 1, and subsequently stabilising to lower levels. Aegolius funereus exhibits generally sparse activity with intermittent detection peaks, notably reaching approximately 450 detections on March 3. Strix uralensis exhibits two significant detection pulses, approximately 400 on February 26 and around 700 on March 4, separated by several days of minimal activity. The observed high variance suggests that a limited number of days with heightened activity disproportionately influence the series, likely reflecting both the short sampling duration and the relatively small dataset. Future research may find that reliance solely on literal detection counts does not accurately represent the actual presence of owls. Instead, more sophisticated post-processing techniques that emphasise the detection of individual owls could provide more reliable insights.

5.3 Implications, limitations, and future work

The findings of this study demonstrate the effectiveness of modern object detection architectures in the realm of passive acoustic monitoring (PAM) of owls. Under consistent training and evaluation conditions, both YOLOv11-l and Faster R-CNN with a ResNet-50 backbone exhibited exceptional detection capabilities, attaining F1 scores of 96.8% and 96.3%, respectively. The confusion matrices (Figures 12 and 14) indicated a lack of interspecific confusions within either architecture, with all remaining errors attributed to misclassification between owl calls and background noise. Collectively, these results imply that fine-tuned object-detection CNNs can deliver the necessary accuracy and robustness for long-term owl PAM.

Reframing species recognition as an object detection problem on time-frequency spectrograms presents several important implications. Each detection is correlated with a designated time-frequency box, which naturally encodes the call onset, duration, and dominant frequency range. This representation provides a more comprehensive foundation for downstream analyses compared to clip-level labels, allowing for the segregation of individual vocalisations within a recording and, with appropriate post-processing, the reconstruction of call sequences and individual encounter histories. The absence of interspecific confusions in the confusion matrices indicates that the detector successfully learned distinctly separated spectral-temporal templates for the three owl species, with remaining classification errors primarily driven by background noise, such as wind, rain, and other species, rather than overlap among target classes.

From a systems perspective, the recorded processing speed of 52.8 seconds per hour of audio (approximately 68 times real-time) demonstrates practical compatibility with affordable autonomous recorders, such as AudioMoth devices. The hardware used in this study allows a single processing node to analyse continuous data streams from multiple recorders efficiently. In theory, this setup could process over a day's worth of audio from a single device in under 30 minutes. This processing margin offers two main opportunities for optimisation. First, it allows for testing more complex models, higher-resolution inputs, longer temporal windows, or small ensemble methods if higher accuracy or more species detection is needed. Second, it opens the possibility of near real-time operation directly on or near the recording hardware. This edge-PAM setup could involve the local computation of spectrograms, running a simplified YOLOv11-l model, and transmitting compact detection summaries (timestamp, species label, confidence, and a rough time-frequency footprint) over low-bandwidth networks, such as LoRaWAN [80], to a nearby base station. These enhancements could significantly cut storage

needs on AudioMoth devices and enable prompt notifications of owl activity in remote forests.

Despite these strengths, several limitations restrict the generalizability of the present findings. The training dataset was relatively small and highly labelled, which facilitates learning but may overestimate performance in more challenging scenarios involving weak, overlapping, or partially obscured calls. The three study species occupy distinct and largely non-overlapping frequency bands (see Figure 9), simplifying their discrimination. Species with more acoustically similar signals or communities containing numerous co-vocalising species are likely to pose greater challenges for the current architecture. Additionally, the detectors were evaluated under ideal conditions where the AudioMoth devices operated as intended. In real-world settings, factors such as extreme cold, moisture, wind, and snow can diminish battery life, alter microphone sensitivity, increase self-noise, and produce transient artefacts that resemble calls, potentially leading to false positives or missed detections. Implementing simple modifications, such as windshields, weather-resistant housings, and mounting away from vibrating structures, along with recording local weather, could help characterise and mitigate these effects. Finally, the reliance on manually annotated bounding boxes limits scalability to communities comprising dozens or hundreds of species. Extending this methodology to more diverse assemblages will require more automated labelling approaches, such as active learning or semi-supervised annotation and training strategies explicitly addressing class imbalance.

As outlined in the introduction, owls play a crucial role in the Finnish boreal forest ecosystem as sensitive indicators of environmental change, with recent concerns focusing on both smaller species, such as Aegolius funereus and Glaucidium passerinum, as well as larger species, including Strix uralensis [5]. These species appear to be contracting in distribution in southern Finland alongside increasing forest fragmentation and the decline of mature forest stands [81]. Accurately demonstrating these trends has been challenging due to reliance on limited and inconsistent monitoring data, resulting in a poor understanding of individual-level abundances. This thesis highlights the potential of automated, spectrogram-based detection methods to address this gap by enabling precise, large-scale call detection, which is essential for advancing future monitoring systems aimed at acoustically identifying individual owls and transitioning from coarse presence indices to robust estimates of population size and site fidelity. Such advancements can provide a more reliable empirical foundation for assessing the impacts of various silvicultural practices, particularly the contrast between clear-cut forestry and continuous-cover management. They can support evidence-based decision-making regarding habitat retention, strategic nest box placement, and the conservation of structurally diverse forest patches, which are essential for

both small and large owl populations. Ultimately, the proposed detection framework serves as an enabling tool for testing the widely held hypothesis that continuous-cover forestry is ecologically preferable for owl populations, transforming anecdotal concerns into quantitatively verifiable conservation insights.

Future work should consequently advance along both methodological and ecological dimensions. Methodologically, it is crucial to expand the dataset by incorporating more sites, seasons, and species, paired with semi-automatic or expert-in-the-loop labelling pipelines. This expansion is essential to evaluate the approach's ability to generalise beyond the relatively favourable conditions encountered in this study. A key design consideration involves deciding whether to deploy a single large multi-species detector or multiple smaller, species-specific models or heads. The results obtained here, in conjunction with previous work on multi-species detection, suggest that utilising a shared backbone with species- or guild-specific detection heads may offer a practical compromise. This approach leverages shared structural features while allowing for more aggressive tuning of rare or high-priority species. Ecologically, future efforts should concentrate on developing more reliable post-processing techniques for individual identification and outlier removal, as well as building coordinated multi-site modelling frameworks that unify detections across stations. Such frameworks would enable tracking of individual movements over time. These advancements aim to improve the reliability of spatial ecology inferences and support evidence-based conservation and management strategies for owl populations in altered landscapes.

6 Conclusion

This thesis evaluated modern object-detection methodologies applied to owl vocalisations within Mel spectrogram representations, with the goal of developing a robust, deployable detector for passive acoustic monitoring. The Faster R-CNN architecture with a ResNet-50 backbone and YOLOv11-l were adapted to spectrogram inputs, utilising strongly labelled time-frequency bounding boxes and adhering to a standardised evaluation protocol. The YOLOv11-l model demonstrated a remarkable F1 score of 96.8% on the target task and exhibited biologically interpretable diel and spatial distribution patterns across different sites. Consequently, it was selected for deployment in large-scale analyses. Although these promising results were achieved using a relatively small, well-annotated dataset and species characterised by distinct frequency signatures, future research should aim to expand species coverage, implement automated labelling procedures, and enhance post-processing techniques. These enhancements are essential for the accurate identification and monitoring of individual specimens across synchronised sites, thereby improving our understanding of owl populations and the impacts of habitat changes within Finland.

References

[1] R. D. Gregory, J. Skorpilova, P. Vorisek, and S. Butler, 'An analysis of trends, uncertainty and species selection shows contrasting trends of widespread forest and farmland birds in Europe', *Ecol Indic*, vol. 103, pp. 676–687, Aug. 2019, doi: 10.1016/J.ECOLIND.2019.04.064.

[2] X. Tang, H. Li, X. Xu, G. Yang, G. Liu, X. Li, and D. Chen, 'Changing land use and its impact on the habitat suitability for wintering Anseriformes in China's Poyang Lake region', *Sci Total Environ*, vol. 557–558, pp. 296–306, 2016, doi: 10.1016/j.scitotenv.2016.03.108.

[3] A. Çelekli, *BIODIVERSITY LOSS: A GLOBAL ISSUE THREATENING ECOLOGICAL BALANCE*. Academy Global Publishing House, 2024. [Online]. Available: https://www.researchgate.net/publication/383240551

[4] J. V. Remsen, 'The importance of continued collecting of bird specimens to ornithology and bird conservation', *Bird Conserv Int*, vol. 5, no. 2–3, pp. 146–180, 1995, doi: 10.1017/S095927090000099X.

[5] D. Baroni, J. Hanzelka, T. Raimondi, M. Gamba, J. E. Brommer, and T. Laaksonen, 'Passive acoustic survey reveals the abundance of a low-density predator and its dependency on mature forests', *Landsc Ecol*, vol. 38, no. 8, pp. 1939–1954, 2023, doi: 10.1007/s10980-023-01667-1.

[6] D. T. Blumstein, D. J. Mennill, P. Clemins, L. Girod, K. Yao, G. Patricelli, J. L. Deppe, A. H. Krakauer, C. Clark, K. A. Cortopassi, S. F. Hanser, B. McCowan, A. M. Ali, and A. N. G. Kirschel, 'Acoustic monitoring in terrestrial environments using microphone arrays: applications, technological considerations and prospectus', *J Appl Ecol*, vol. 48, no. 3, pp. 758–767, 2011, doi: 10.1111/j.1365-2664.2011.01993.x.

[7] T. A. Marques, L. Thomas, S. W. Martin, D. K. Mellinger, J. A. Ward, D. J. Moretti, D. Harris, and P. L. Tyack, 'Estimating animal population density using passive acoustics', *Biol Rev Camb Philos Soc*, vol. 88, no. 2, pp. 287–309, 2013, doi: 10.1111/brv.12001.

[8] A. Farina, N. Pieretti, and L. Piccioli, 'The soundscape methodology for long-term bird monitoring: A Mediterranean Europe case-study', *Ecol Inform*, vol. 6, no. 6, pp. 354–363, 2011, doi: 10.1016/j.ecoinf.2011.07.004.

[9] K. J. Piczak, 'Environmental sound classification with convolutional neural networks', in *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, 2015, pp. 1–6. doi: 10.1109/MLSP.2015.7324337.

[10] K. Choi, G. Fazekas, M. Sandler, and K. Cho, 'A Comparison of Audio Signal Preprocessing Methods for Deep Neural Networks on Music Tagging', in *2018 26th European Signal Processing Conference (EUSIPCO)*, 2018, pp. 1870–1874. doi: 10.23919/EUSIPCO.2018.8553106.

[11] K. Simonyan and A. Zisserman, 'Very deep convolutional networks for large-scale image recognition', in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015. doi: 10.48550/arXiv.1409.1556.

[12] K. He, X. Zhang, S. Ren, and J. Sun, 'Deep Residual Learning for Image Recognition', in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[13] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, 'You Only Look Once: Unified, Real-Time Object Detection', in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788. doi: 10.1109/CVPR.2016.91.

[14] R. Shrestha, C. Glackin, J. Wall, and N. Cannings, 'Bird Audio Diarization with Faster R-CNN', in Artificial Neural Networks and Machine Learning – ICANN 2021, vol. 12891. Springer, 2021, pp. 415–426. doi: 10.1007/978-3-030-86362-3_34.

[15] A. Márquez-Rodríguez, M. Á. Mohedano-Munoz, M. J. Marín-Jiménez, E. Santamaría-García, G. Bastianelli, P. Jordano, and I. Mendoza, 'A bird song detector for improving bird identification through deep learning: A case study from Doñana', *Ecol Inform*, vol. 90, p. 103254, 2025, doi: 10.1016/j.ecoinf.2025.103254.

[16] S. Kahl, C. M. Wood, M. Eibl, and H. Klinck, 'BirdNET: A deep learning solution for avian diversity monitoring', *Ecol Inform*, vol. 61, p. 101236, 2021, doi: 10.1016/j.ecoinf.2021.101236.

[17] Z. J. Ruff, D. B. Lesmeister, L. S. Duchac, B. K. Padmaraju, C. M. Sullivan, N. Pettorelli, and V. Lecours, 'Automated identification of avian vocalizations with deep convolutional neural networks', *Remote Sens Ecol Conserv*, vol. 6, no. 1, pp. 79–92, 2020, doi: 10.1002/rse2.125.

[18] S. Ren, 'Faster R-CNN: Towards real-time object detection with region proposal networks', *Adv Neural Inf Process Syst*, vol. 2015-January, pp. 91–99, 2015.

[19] S. J. White, R. E. C. White, and W. H. Thorpe, 'Acoustic Basis for Individual Recognition by Voice in the Gannet', *Nature*, vol. 225, no. 5238, pp. 1156–1158, 1970, doi: 10.1038/2251156a0.

[20] T. Lengagne, 'Intra-syllabic acoustic signatures used by the king penguin in parent-chick recognition: An experimental approach', *Journal of Experimental Biology*, vol. 204, no. 4, pp. 663–672, 2001.

[21] Z. Zhao, S. Zhang, Z. Xu, K. Bellisario, N. Dai, H. Omrani, and B. C. Pijanowski, 'Automated bird acoustic event detection and robust species classification', *Ecol Inform*, vol. 39, pp. 99–108, 2017, doi: 10.1016/j.ecoinf.2017.04.003.

[22] S. Chu, S. Narayanan, and C.-C. J. Kuo, 'Environmental Sound Recognition With Time–Frequency Audio Features', *IEEE Trans Audio Speech Lang*

Process, vol. 17, no. 6, pp. 1142–1158, 2009, doi: 10.1109/TASL.2009.2017438.

[23] Z. K. Abdul and A. K. Al-Talabani, 'Mel Frequency Cepstral Coefficient and its Applications: A Review', *IEEE Access*, vol. 10, pp. 122136–122158, 2022, doi: 10.1109/ACCESS.2022.3223444.

[24] J. Zhao, X. Mao, and L. Chen, 'Speech emotion recognition using deep 1D & 2D CNN LSTM networks', *Biomed Signal Process Control*, vol. 47, pp. 312–323, 2019, doi: 10.1016/j.bspc.2018.08.035.

[25] M. T. García-Ordás, S. Rubio-Martín, J. A. Benítez-Andrades, H. Alaiz-Moretón, and I. García-Rodríguez, 'Multispecies bird sound recognition using a fully convolutional neural network', *Applied intelligence.*, vol. 53, no. 20, pp. 23287–23300, 2023, doi: 10.1007/s10489-023-04704-3.

[26] A. A. Hidayat, T. W. Cenggoro, and B. Pardamean, 'Convolutional Neural Networks for Scops Owl Sound Classification', *Procedia Comput Sci*, vol. 179, pp. 81–87, 2021, doi: 10.1016/j.procs.2020.12.010.

[27] L. Fanioudakis and I. Potamitis, 'Deep Networks tag the location of bird vocalisations on audio spectrograms', 2017. [Online]. Available: http://www.xeno-canto.org/

[28] D. Stowell, M. D. Wood, H. Pamuła, Y. Stylianou, H. Glotin, and D. Orme, 'Automatic acoustic detection of birds through deep learning: The first Bird Audio Detection challenge', *Methods in ecology and evolution /*, vol. 10, no. 3, pp. 368–380, 2019, doi: 10.1111/2041-210X.13103.

[29] S. J. Pan and Q. Yang, 'A Survey on Transfer Learning', *IEEE Trans Knowl Data Eng*, vol. 22, no. 10, pp. 1345–1359, 2010, doi: 10.1109/TKDE.2009.191.

[30] A. Krizhevsky, 'ImageNet classification with deep convolutional neural networks', *Adv Neural Inf Process Syst*, vol. 2, pp. 1097–1105, 2012.

[31] E. Fanioudakis, M. Geismar, and I. Potamitis, 'Mosquito wingbeat analysis and classification using deep learning', in *2018 26th European Signal Processing Conference (EUSIPCO)*, 2018, pp. 2410–2414. doi: 10.23919/EUSIPCO.2018.8553542.

[32] M. Zhong, J. LeBien, M. Campos-Cerqueira, R. Dodhia, J. Lavista Ferres, J. P. Velev, and T. M. Aide, 'Multispecies bioacoustic classification using transfer learning of deep convolutional neural networks with pseudo-labeling', *Applied Acoustics*, vol. 166, Sep. 2020, doi: 10.1016/j.apacoust.2020.107375.

[33] D. Konovalov and M. Sankupellay, 'Bird Call Recognition using Deep Convolutional Neural Network, ResNet-50', 2018, doi: 10.13140/RG.2.2.31865.31847.

[34] E. J. Henri and Z. Mungloo-Dilmohamud, 'A deep transfer learning model for the identification of bird songs: A case study for mauritius', in

International Conference on Electrical, Computer, Communications and Mechatronics Engineering, ICECCME 2021, Institute of Electrical and Electronics Engineers Inc., Oct. 2021. doi: 10.1109/ICECCME52200.2021.9590917.

[35] M. Schuster and K. Paliwal, 'Bidirectional recurrent neural networks', *IEEE Trans Signal Process*, vol. 45, no. 11, pp. 2673–2681, 1997, doi: 10.1109/78.650093.

[36] S. Hochreiter and J. Schmidhuber, 'Long Short-Term Memory', *Neural Comput*, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.

[37] R. Dey and F. M. Salem, 'Gate-variants of Gated Recurrent Unit (GRU) neural networks', in *2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS)*, 2017, pp. 1597–1600. doi: 10.1109/MWSCAS.2017.8053243.

[38] A. Noumida and R. Rajan, 'Multi-label bird species classification from audio recordings using attention framework', *Applied acoustics.*, vol. 197, p. 108901, 2022, doi: 10.1016/j.apacoust.2022.108901.

[39] A. Saad, M. M. Ahmad Zabidi, I. S. Kamil, and U. U. Sheikh, 'A Novel Deep Learning Approach for Classification of Bird Sound Using Mel Frequency Cepstral Coefficients', *Iraqi journal for computer science and mathematics.*, vol. 5, no. 3, pp. 660–670, 2024, doi: 10.52866/ijcsm.2024.05.03.040.

[40] C. Li, B. Zhang, H. Hu, and J. Dai, 'Enhanced Bird Detection from Low-Resolution Aerial Image Using Deep Neural Networks', *Neural Process Lett*, vol. 49, no. 3, pp. 1021–1039, 2019, doi: 10.1007/s11063-018-9871-z.

[41] Z. Yang, J. Wang, J. Shen, Z. Ling, D. Zhang, and C. Jing, 'Birdsong recognition method based on Mel spectrogram and YOLO-BEM', in *2024 7th International Conference on Computer Information Science and Application Technology, CISAT 2024*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 401–405. doi: 10.1109/CISAT62382.2024.10695385.

[42] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, 'Feature Pyramid Networks for Object Detection', in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 936–944. doi: 10.1109/CVPR.2017.106.

[43] A. Neubeck and L. Van Gool, 'Efficient Non-Maximum Suppression', in *18th International Conference on Pattern Recognition (ICPR'06)*, 2006, pp. 850–855. doi: 10.1109/ICPR.2006.479.

[44] M. L. Ali and Z. Zhang, 'The YOLO Framework: A Comprehensive Review of Evolution, Applications, and Benchmarks in Object Detection', *Computers.*, vol. 13, no. 12, p. 336, 2024, doi: 10.3390/computers13120336.

[45] E. Reswara, S. Suakanto, and S. A. Putra, 'Comparison of Object Detection Algorithm using YOLO vs Faster R-CNN : A Systematic Literature Review', in *Proceedings of the 2023 6th International Conference on Big Data*

Technologies, in ICBDT '23. New York, NY, USA: Association for Computing Machinery, 2023, pp. 419–424. doi: 10.1145/3627377.3627443.

[46] Y. Lechevallier, G. Saporta, and L. Bottou, *Large-Scale Machine Learning with Stochastic Gradient Descent*. 2010. doi: 10.1007/978-3-7908-2604-3_16.

[47] D. P. Kingma and J. L. Ba, 'Adam: A method for stochastic optimization', in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015. doi: 10.48550/arXiv.1412.6980.

[48] Z. Zhang, 'Improved Adam Optimizer for Deep Neural Networks', in *2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS)*, 2018, pp. 1–2. doi: 10.1109/IWQoS.2018.8624183.

[49] I. Loshchilov and F. Hutter, 'Decoupled weight decay regularization', in *7th International Conference on Learning Representations Iclr 2019*, 2019. doi: 10.48550/arXiv.1711.05101.

[50] R. G. Baraniuk, 'Compressive Sensing [Lecture Notes]', *IEEE Signal Process Mag*, vol. 24, no. 4, pp. 118–121, 2007, doi: 10.1109/MSP.2007.4286571.

[51] S. Mallat, *A Wavelet Tour of Signal Processing : The Sparse Way*. Chantilly, UNITED STATES: Elsevier Science & Technology, 2008. [Online]. Available: http://ebookcentral.proquest.com/lib/aalto-ebooks/detail.action?docID=413871

[52] T. Bradfer-Lawrence, Z. Buřivalová, and D. H. Dent, 'Deriving meaning from acoustic data in hyper-diverse ecosystems', May 01, 2025, *Elsevier Ltd*. doi: 10.1016/j.tree.2025.03.004.

[53] Tom Bäckström, 'Introduction to Speech Processing (2nd ed.)'. Accessed: Oct. 01, 2025. [Online]. Available: https://speechprocessingbook.aalto.fi/index.html#introduction-to-speech-processing

[54] F. Han and J. Hu, 'Distribution and strength of sound in the human head', in *Proceedings of the 2015 Symposium on Piezoelectricity, Acoustic Waves, and Device Applications (SPAWDA 2015) : Oct. 30-Nov. 2, 2015, Shandong University /*, in 2015 Symposium on Piezoelectricity, Acoustic Waves, and Device Applications (SPAWDA). Piscataway, NJ : IEEE, 2015, pp. 96–105. doi: 10.1109/SPAWDA.2015.7364450.

[55] S. S. Stevens, J. Volkmann, and E. B. Newman, 'A Scale for the Measurement of the Psychological Magnitude Pitch', *J Acoust Soc Am*, vol. 8, no. 3, pp. 185–190, 1937, doi: 10.1121/1.1915893.

[56] D. Issa, M. Fatih Demirci, and A. Yazici, 'Speech emotion recognition with deep convolutional neural networks', *Biomed Signal Process Control*, vol. 59, p. 101894, 2020, doi: 10.1016/j.bspc.2020.101894.

[57] S. Umesh, L. Cohen, and D. Nelson, 'Fitting the Mel scale', in *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing*.

Proceedings. ICASSP99 (Cat. No.99CH36258), 1999, pp. 217–220 vol.1. doi: 10.1109/ICASSP.1999.758101.

[58] G. van Rossum, 'Python tutorial', 1995.

[59] A. Paszke, 'PyTorch: An imperative style, high-performance deep learning library', *Adv Neural Inf Process Syst*, vol. 32, 2019.

[60] G. Jocher, J. Qiu, and A. Chaurasia, 'Ultralytics YOLO (Version 8.0.0)', 2023, *Ultralytics*. Accessed: Oct. 25, 2025. [Online]. Available: https://github.com/ultralytics/ultralytics

[61] B. Mcfee, C. Raffel, D. Liang, D. P. W. Ellis, M. Mcvicar, E. Battenberg, and O. Nieto, 'librosa: Audio and Music Signal Analysis in Python', 2015. [Online]. Available: https://www.youtube.com/watch?v=MhOdbtPhbLU

[62] A. Robitaille, 'Folium: Python Data. Leaflet.js Maps', 203AD. Accessed: Oct. 26, 2025. [Online]. Available: https://python-visualization.github.io/folium/

[63] S. Kivinen and P. Nummi, 'Immediate facilitation and engineering legacy of beavers: 54 years of patch dynamics in a boreal landscape', *Sci Total Environ*, vol. 999, p. 180341, 2025, doi: 10.1016/j.scitotenv.2025.180341.

[64] OpenStreetMap contributors, 'Planet dump retrieved from https://planet.osm.org ', 2017. Accessed: Nov. 05, 2025. [Online]. Available: https://www.openstreetmap.org/

[65] 'AudioMoth Dev 1.0.1 Datasheet'. Accessed: Nov. 04, 2025. [Online]. Available: http://github.com/OpenAcousticDevices/Datasheets/blob/main/AudioMoth_Dev_1_0_1_Datasheet/AudioMoth_Dev_1_0_1_Datasheet.pdf

[66] J. E. Brommer and H. Kolunen, 'Reproduction and Survival in a Variable Environment: Ural Owls (Strix uralensis) and the Three-Year Vole Cycle', *Auk*, vol. 119, no. 2, pp. 544–550, 2002, doi: 10.2307/4089904.

[67] G. A. Sonerud, 'Nest Hole Shift in Tengmalm's Owl Aegolius funereus as Defence Against Nest Predation Involving Long-Term Memory in the Predator', *J Anim Ecol*, vol. 54, no. 1, pp. 179–192, 1985, doi: 10.2307/4629.

[68] R. K. Heikkinen, M. Luoto, R. Virkkala, R. G. Pearson, and J. Körber, 'Biotic interactions improve prediction of boreal bird distributions at macro-scales', *Global ecology & biogeography.*, vol. 16, no. 6, pp. 754–763, 2007, doi: 10.1111/j.1466-8238.2007.00345.x.

[69] G. Andrey, 'Ural Owl'. Accessed: Oct. 26, 2025. [Online]. Available: https://www.flickr.com/photos/184846830@N08/53638675895/

[70] J. S. Charles, 'Eurasian pygmy owl (Glaucidium passerinum)'. Accessed: Oct. 26, 2025. [Online]. Available: www.flickr.com/photos/charlesjsharp/52889724684

[71] S. Vil, 'Raufußkauz', 2014, Accessed: Oct. 26, 2025. [Online]. Available: https://www.flickr.com/photos/vil_sandi/15431521098

[72] E. E. Waddell, J. H. Rasmussen, and A. Širović, 'Applying artificial intelligence methods to detect and classify fish calls from the northern gulf of mexico', *J Mar Sci Eng*, vol. 9, no. 10, Oct. 2021, doi: 10.3390/jmse9101128.

[73] Y. Gong and C. Poellabauer, 'Impact of Aliasing on Deep CNN-Based End-to-End Acoustic Models', in *Interspeech 2018*, in Interspeech 2018, vol. 2018- September. ISCA, 2018, pp. 2698–2702. doi: 10.21437/Interspeech.2018-1371.

[74] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, 'Microsoft COCO: Common Objects in Context', in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., Cham: Springer International Publishing, 2014, pp. 740–755.

[75] T. M. Aide, C. Corrada-Bravo, M. Campos-Cerqueira, C. Milan, G. Vega, and R. Alvarez, 'Real-time bioacoustics monitoring and automated species identification', *PeerJ*, vol. 2013, no. 1, p. e103, 2013, doi: 10.7717/peerj.103.

[76] J. Kaur, S. Rani, A. Sharma, and A. Dogra, 'Enhanced Bird Species Identification using ResNet-50: A Deep Learning Framework for High-Performance Classification', in *3rd International Conference on Automation, Computing and Renewable Systems, ICACRS 2024 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 821–826. doi: 10.1109/ICACRS62842.2024.10841723.

[77] A. Kumar, R. Shen, S. Bubeck, and S. Gunasekar, 'HOW TO FINE-TUNE VISION MODELS WITH SGD', in *12th International Conference on Learning Representations Iclr 2024*, Hybrid, Vienna, 2024. doi: 10.48550/arXiv.2211.09359.

[78] R. Girshick, J. Donahue, T. Darrell, and J. Malik, 'Region-Based Convolutional Networks for Accurate Object Detection and Segmentation', *IEEE Trans Pattern Anal Mach Intell*, vol. 38, no. 1, pp. 142–158, 2016, doi: 10.1109/TPAMI.2015.2437384.

[79] R. Zeiler Matthew D. and Fergus, 'Visualizing and Understanding Convolutional Networks', in *Computer Vision – ECCV 2014*, T. and S. B. and T. T. Fleet David and Pajdla, Ed., Cham: Springer International Publishing, 2014, pp. 818–833.

[80] F. Adelantado, X. Vilajosana, P. Tuset-Peiro, B. Martinez, J. Melia-Segui, and T. Watteyne, 'Understanding the Limits of LoRaWAN', *IEEE Communications Magazine*, vol. 55, no. 9, pp. 34–40, 2017, doi: 10.1109/MCOM.2017.1600613.

[81] T. Solonen, J. Karhunen, J. A. Kekkonen, H. Kolunen, and H. Pietiäinen, 'Tawny owl prey remains indicate differences in the dynamics of coastal and inland vole populations in southern Finland', *Popul Ecol*, vol. 58, no. 4, pp. 557–565, 2016, doi: https://doi.org/10.1007/s10144-016-0556-z.

Appendix A - Site Information

Site	Latitude (decimal degrees)	Longitude (decimal degrees)	Year of collection	Part of the dataset
Ko1	61.2410531	25.067693	2023	1
Ko2	61.241818	25.0727541	2023	1
Ko3	61.2439189	25.067423	2023	1
Ko4	61.2461766	25.070338	2023	1
Ko5	61.2455033	25.0623183	2023	1
Ko6	61.2439833	25.0578524	2023	1
Ko7	61.2409418	25.0610892	2023	1
Ko8	61.2388151	25.0590559	2023	1
Ko9	61.2413832	25.0554952	2023	-
Ko10	61.2431093	25.0498149	2023	1
Ko11	61.2469928	25.0474222	2023	-
Ko12	61.2485181	25.062956	2023	1
Ko13	61.2512107	25.0611773	2023	1
Ko14	61.25110521	25.0553211	2023	1
Ru1	61.2186314	25.0225111	2023	1
Ru2	61.2156593	25.0266539	2023	-
Ru3	61.2132648	25.0336125	2023	-
Ru4	61.2106711	25.0427933	2023	-
Ru5	61.2102304	25.0511404	2023	-
Ru6	61.2137975	25.0543279	2023	-
Ru7	61.2169274	25.0720651	2023	-
Ru8	61.2201582	25.0696598	2023	-
Ru9	61.2221489	25.062213	2023	-
Ru10	61.2201343	25.0582141	2023	-
Ru11	61.22299	25.0539438	2023	-
Ru12	61.2261165	25.0595407	2023	1
Ru13	61.2284144	25.0756976	2023	1
Ru14	61.2307798	25.0785988	2023	1
Ru15	61.2318646	25.092968	2023	1
Ru16	61.2271884	25.0914233	2023	1
Su1	61.2007954	25.1926092	2023	-
Su2	61.2018484	25.199325	2023	-
Su3	61.2022992	25.20514	2023	-
Su4	61.2048852	25.2081947	2023	-
Su5	61.2051477	25.2141905	2023	1
Su6	61.2024666	25.18974233	2023	1
Su7	61.2039503	25.1961617	2023	-
Su8	61.204904	25.2010175	2023	1
Su9	61.2072908	25.2087926	2023	-
Su10	61.2055966	25.1886066	2023	-
Su11	61.2073434	25.1976393	2023	-

Su12	61.2075382	25.204293	2023	-
Su13	61.2083988	25.1860443	2023	1
Su14	61.2079436	25.1922614	2023	-
Su15	61.2100072	25.1969532	2023	-
Su16	61.2107202	25.187428	2023	1
Su17	61.2107619	25.1937304	2023	-
Su18	61.2123939	25.2017362	2023	1
Su19	61.2120024	25.182855	2023	-
Su20	61.212317	25.1852958	2023	-
Su21	61.2139991	25.1924089	2023	-
Su22	61.2159622	25.1983163	2023	-
Su23	61.2171768	25.2104606	2023	-
Su24	61.2148628	25.1827748	2023	1
Su25	61.21709	25.1899133	2023	-
Su26	61.2191462	25.1934956	2023	-
Su27	61.2184805	25.1998309	2023	-
Su28	61.2021242	25.1761491	2023	1
Su29	61.2037908	25.1712011	2023	1
Ha1	61.2334588	25.1894532	2023	1
Ha2	61.237436	25.1896591	2023	-
Ha3	61.2392864	25.1938787	2023	-
Ha6	61.2385583	25.1780913	2023	-
Ha7	61.2417736	25.1774469	2023	1
Tuohi01	61.2028574	25.1691001	2025	2
Tuohi02	61.2026866	25.175014	2025	1
Tuohi03	61.2018219	25.1773365	2025	2
Koti01	61.2438849	25.0481515	2025	1
Koti02	61.2404047	25.0682795	2025	1
Koti03	61.2403733	25.0627688	2025	1
Koti04	61.2433789	25.0722626	2025	1
Haara01	61.236977	25.1902316	2025	1
Haara02	61.2382957	25.1825962	2025	1
Nime01	61.2277872	25.1920973	2025	1
Susi01	61.2213037	25.2003352	2025	1
Ruuhi01	61.2237336	25.0641203	2025	1
Upm06	60.98374	24.80154	2025	1
Upm07	60.98561	24.79254	2025	1
Upm08	60.98506	24.78418	2025	1
Upm09	60.98718	24.77674	2025	2
Upm10	60.99153	24.78044	2025	2
Upm11	60.99477	24.78201	2025	2
Upm12	60.99206	24.77167	2025	2
Upm13	60.99792	24.75188	2025	2
Upm14	60.99663	24.74429	2025	2
Upm15	60.99049	24.74496	2025	1
Upm21	60.98257	24.77696	2025	2

Upm22	60.97774	24.77694	2025	2
Upm23	60.97488	24.77173	2025	2
Upm24	60.97084	24.76816	2025	1
Upm25	60.96793	24.7599	2025	2

Table 1: Raw information of each site. Name, latitude, longitude, year and the dataset it is used in