Acoustics Australia
https://doi.org/10.1007/s40857-026-00395-1

RESEARCH

# Cross-Model Confusion Mapping: False Alarm Minimisation in Possum Call Detection Using Deep Learning

**Akbar Ghobakhlou**[1] · **Mohammad A. Barzegartabrizi**[1] · **Edmund Lai**[1]

Received: 14 April 2026 / Accepted: 10 June 2026


**Abstract**
The common brushtail possum is a major invasive species in New Zealand, damaging native ecosystems and agricultural systems and is a target of the national Predator Free 2050 eradication programme. Reliable, low-cost monitoring is critical to this programme, and acoustic detection offers a scalable alternative to traditional trapping and camera-based methods. The bioacoustics approach to automated brushtail possum detection using deep learning models suffers from excessive false alarms. At the same time, the high computational demands of the most accurate models make them unsuitable for resource-limited edge devices. In this paper, a targeted hard-negative mining method, Cross-Model Confusion Mapping (CMCM), is proposed to overcome these problems. It uses the error profile of a pre-trained model to identify bird calls most frequently misclassified as possum vocalisations. These bird calls are then relabelled and added to the training data. Three CMCM audiosets were generated and compared with size-matched sets assembled through untargeted sampling. CNNs with 3, 5, 7, 10, and 13 layers were trained using identical augmentation pipelines. Experiments showed that CMCM-trained models maintain accuracy while significantly reducing false alarms. In particular, the 10-layer CMCM-trained CNN achieved the most reliable detection with zero false positives. The inference times are orders of magnitude faster, with substantially lower computational demand, than the state-of-the-art Audio Spectrogram Transformer. These findings suggest that CMCM enables accurate, low-latency possum detection suitable for real-time deployment on edge devices. The method is generalisable to other bioacoustic detection tasks where false positives arise from acoustically similar non-target species.

**Keywords** Bioacoustics · Deep learning · Possum detection · False positive reduction · Hard-negative mining · Edge deployment

## 1 Introduction

New Zealand faces an ecological challenge due to invasive mammal species, particularly the common brushtail possum (Trichosurus vulpecula) [1]. The brushtail possum was introduced to New Zealand in 1858 for fur harvesting and has since become widespread and abundant [2]. They have had catastrophic impacts on native flora and fauna, outcompeting indigenous species [3] and spreading bovine tuberculosis [4]. Possums can damage agricultural crops, spread disease, and cause major changes in native ecosystems [5]. Predator Free 2050, launched by the New Zealand government in July 2016, aims to eradicate introduced predators such as possums by mid-century [6].

Targeted eradication efforts depend on the ability to estimate the distribution and abundance of possums across the landscape [6] at various stages of eradication. During the "mop-up" phase, detecting the last remaining individuals is particularly challenging due to low numbers. However, it is important that they are detected and removed to prevent re-establishment [7]. The sensitivity of traditional possum monitoring techniques such as trapping lines, chew cards [8], and wax tracking tags [7], declines substantially as the numbers decrease [7]. Camera traps (motion-activated wildlife cameras) [9–11] can collect photographic evidence of possums, and the manual review burden has been substantially reduced by automated image classification. However, camera traps sample a narrow field of view at short detection distance, whereas acoustic recorders sample a 360-degree volume with a larger detection radius, making them better

✉ Mohammad A. Barzegartabrizi
amin.barzegar@aut.ac.nz

[1] Data Science and AI Department Level 3, Auckland University of Technology, WZ Building, 6 St Paul Street, Auckland Central, Auckland 1010, New Zealand

Published online: 24 June 2026

suited to sparse-population monitoring such as the mop-up phase of eradication. The trade-off is that acoustic detection is restricted to soniferous species.

### 1.1 Related Work

Deep learning models are the most accurate machine learning methods for species classification. Models designed specifically for audio classification include Audio Spectrogram Transformer (AST) [12], MobileNet [13] and ResNET [14]. More recently, a fine-tuned AST model was utilised to classify possum vocalisations [15]. It was pre-trained on 527 audio classes, and then fine-tuned with a binary classifier head to detect possum vocalisation and achieved a state-of-the-art F1 score of 0.983 on a validation set of 500 labelled samples. However, McEwen et al. [15] noted that certain vocalisations were frequently misclassified, while bird songs contributed to 30% of false positives.

### 1.2 Contributions

In this paper, we propose Cross-Model Confusion Mapping (CMCM), a targeted hard-negative mining method that leverages an external pre-trained classifier to systematically identify the acoustic sources most likely to trigger false positives. Unlike existing iterative retraining approaches, CMCM is single-pass and uses a probe classifier whose label set is disjoint from the target class. The probe's most-frequent confusions act as a class-level signal that directly identifies acoustic neighbours of the target without requiring an initial target-class detector to be partially trained first. We evaluate CMCM by training five CNN architectures of varying depth on both targeted and untargeted datasets, and we benchmark all models against a fine-tuned Audio Spectrogram Transformer on a real-world evaluation set. The remainder of this paper is organised as follows. Section 2 reviews prior work on reducing false positives in bioacoustic classification. Section 3 describes the CMCM method. The experimental design is presented in Sect. 4, followed by results and discussion in Sect. 5. Conclusions and future work are given in Sect. 6.

## 2 Dealing with False Positives

A major challenge in using deep learning models for bioacoustics possum detection is the high rate of false alarms (false positives) resulting from a broad range of bird calls which the model has not been trained to detect. Several studies in the field of bioacoustics aimed at maximising precision by deliberately incorporating hard negatives into training data to boost the precision in classifiers. The term "hard negatives" is used to refer to the inputs that are difficult for a model to distinguish from the true-positive inputs, resulting in false positives.

Merchan et al. [16] trained an initial binary CNN classifier on manatee recordings. True manatee calls together with this classifier's false triggers were then used as positive and negative training samples for a second model. They showed that this second model achieved higher precision. LeBien et al. [17] used the same approach to train a classifier for bird and frog calls in tropical soundscapes. In the identification of humpback whales, Allen et al. [18] processed 187,000 h of recordings using an initial CNN classifier. The hard negatives of this classifier were relabelled as noise. They were then added to the highest scoring audio clips that were not classified as true positives as training data for the next round. This process is repeated multiple times until the precision stabilises. Their final model achieved 97% precision.

These three studies augment the original training data by harvesting the model's misclassifications. Detection thresholds are deliberately relaxed to capture borderline false positives. Manually confirmed hard negatives are then used to retrain a secondary classifier for these difficult cases. Applying this approach in general, however, requires exposing the model to every major sound that is causing confusion. If such sound sources are diverse, then this process will be very challenging and time-consuming.

## 3 Cross-Model Confusion Mapping

In the context of detecting possums through a binary classifier, a hard negative could be a bird call that the model falsely identifies as a possum call. Given the large diversity of avian species present in the wild, it will be quite impractical to use the approach outlined in the previous section. To overcome this problem, we use a pre-trained model, which will be referred to as the probe model, to relabel the initial dataset. In the context of possum call detection, this relabelling process is illustrated in Fig. 1. This process converts an untargeted dataset into a targeted dataset by using the probe model. The targeted dataset is then used to train the possum call classifier. The aim is to relabel the sound segments that most likely will trigger false positives.

A good probe model to use for possum call detection is BirdNet [20]. This is because bird calls are often misclassified as possum. Since "possum" is absent from BirdNet's label set, the model classifies each possum sound clip to one of the bird species it judges to be most acoustically similar. Audio from these implicated bird species is then collected, manually verified, and incorporated as hard negatives in the training set. We call our proposed method Cross-Model Confusion Mapping (CMCM) since the systematically misclassified labels constitute a confusion map. Although BirdNET is used here

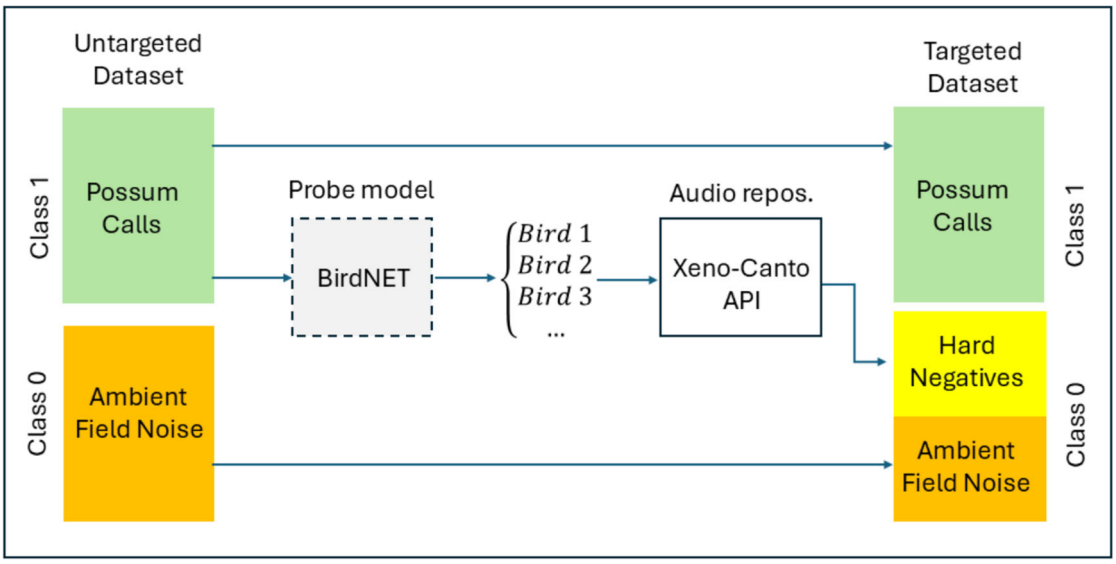


**Fig. 1** Cross-Model Confusion Mapping for hard-negative mining

as the probe model, any pre-trained classifier whose label set excludes the target species can serve the same role.

### 3.1 Positioning of CMCM

CMCM differs from the iterative hard-negative mining approaches discussed above in three structural respects. External probe: CMCM uses a fixed pre-trained classifier (the probe) that is never updated and has no prior exposure to the target species; the iterative train–mine–retrain loop is therefore eliminated. Disjoint label spaces: the probe's label set is deliberately constructed to exclude the target class, so every probe prediction on target audio is, by construction, a mapping to the probe's acoustically nearest known class. The aggregated pattern of these forced misclassifications constitutes the confusion map. Class-level mining: the output of CMCM is a ranked list of confusable classes (here, bird species) rather than a list of individual false-positive clips, making bulk data collection from public repositories far cheaper than the per-clip verification required by standard hard-negative mining.

CMCM is also distinct from active learning, which selects which samples a human should label next using uncertainty or diversity criteria from the model under training. CMCM neither performs sample selection nor requires a human in the loop during mining. The two methods are complementary: probe-model confusion scores could serve as a class-level acquisition signal in an active-learning pipeline.

The general idea of using a foreign classifier's prediction structure to guide training has appeared in adjacent fields, including the use of attribute vectors from foreign classes as hard negatives in zero-shot classification [19], and teacher-model confusions as hard-negative seeds in knowledge-distillation-style image retrieval, but, to our knowledge, has not previously been formalised and evaluated for bioacoustic false-positive reduction.

## 4 Experimental Design

In this study, BirdNET v2.4, which has been trained on 6,147 bird species, serves as the probe model. Bird species that BirdNET repeatedly confuses with possum calls are chosen to populate CMCM-curated datasets. They will be compared with datasets containing only randomly selected forest noise as the negative class to isolate the effect of targeted hard-negative sampling. The confusing bird calls, or hard negatives, were obtained from Xeno-Canto [21].

### 4.1 Audio Sources

A total of seven audio sets are used in the experiments designed to evaluate the effectiveness of the CMCM training approach. These datasets are prepared from three data sources. The first source consists of field recordings of possum calls and ambient sounds obtained from Kaggle [22]. A total of 3500 samples are included, of which 1237 contain possum calls. The second source comprises non-possum audio provided by McEwen B et al. [21], consisting of 1698 five-second ambient field recordings. Ambient sounds such as wind, rain, distant animal calls, and possum movement noises are included, forming part of the non-possum audio segments of the datasets. The third source is bird audio data obtained from Xeno-canto [21], a large open-access repository of bird vocalisations. A total of 200 recordings for each bird species of interest are obtained. In the datasets, each

audio clip is standardised to 5 s in length at a sampling rate of 16 kHz. All audio clips are annotated using custom-built annotation software [23].

Data augmentation is used to compensate for the limited number of audio clips available. Augmented audio data are generated from every 5-s audio clip using a combination of five methods in two groups. The first group consists of time shifting and volume scaling, which do not alter the frequency profile of the audio. The second involves altering the frequency profile through time stretching, pitch shifting, and adding Gaussian noise. Each method is parameterised by hyperparameter $\alpha$ as follows:

- **Pitch-shift (PS):** $\pm\ \alpha$ semitones
- **Time-stretch (TS):** with rate $U(1 - \alpha,\ 1 + \alpha)$; the result was trimmed/padded to N samples.
- **Gaussian noise (GN):** zero-mean Gaussian noise, $\sigma = \alpha$.
- **Time-shift (TSFT):** circular roll of the waveform by $U(1 - \alpha, 1 + \alpha)$. samples (N = clip length).
- **Volume scaling (VS):** multiplicative gain $U(1 - \alpha, 1 + \alpha)$.

where $U(a, b)$ denotes uniform distribution between $a$ and $b$, inclusive. The values of the parameters used are shown in Table 1.

#### 4.1.1 Audiosets

Seven audiosets are prepared and used in the experiments, and their composition is summarised in Table 2. Audioset 0 (AST Set) is defined as an exact copy of the public Kaggle dataset released by McEwen et al. [15]. Of the 3500 audio clips, 3000 are allocated for training and 500 for testing. Within the training set, 1,130 clips are identified as possum vocalisations, resulting in an imbalanced class distribution that is also reflected in the test set.

Audiosets 1 to 6 can be separated into two distinct categories: Untargeted (audiosets 1 to 3) and Targeted through CMCM (audiosets 4–6). In the untargeted audiosets, the non-possum audio segments (Class 0) are selected randomly. They are called "untargeted" due to random selection from this class. In the targeted audiosets, 25% of the non-possum audio segments (Class 0) were replaced with the hard negatives mined using the CMCM method. This ratio (25%) is a rough estimation of the ratio of bird calls to ambient sounds in New Zealand forests.

Apart from these seven audiosets, a separate audioset entirely unseen by the models was prepared. Its purpose is to assess a classifier's precision, recall, and false-positive behaviour in unconstrained acoustic environments. This audioset is referred to as "Real-World" and is extracted from three sources of recordings. The first one is a one-hour night-time soundscape from New Zealand native forest released by the Department of Conservation [24]. It contains night-time bird calls with no possum vocalisations. The second is from a 41 s community video featuring nine clearly audible possum calls recorded at close range (*Brushtail possum making weird noises*, 2016). The third is a 90 s community recording that included seven possum calls interspersed with rustling, insect noise, bird calls and human vocals [26]. This real-world audioset was deliberately chosen to include diverse acoustic conditions absent from the training data, providing a stringent test of each model's ability to generalise.

### 4.2 CNN Architectures and Training

Five convolutional neural network (CNN) models, L3, L5, L7, L10, and L13, were trained using PyTorch 3.12 [27], where the suffix indicates the number of convolutional blocks. The input to every model was a single-channel, 64-bin Mel-spectrogram (a perceptually scaled spectral representation of the audio) derived from 5-s, 16 kHz audio (512-sample hop, 0–8 kHz range). Figure 2 illustrates the architecture of the 5-layer convolutional neural network (CNN) used in this study. This structure is representative of the other CNN variants (3, 7, 10, and 13 layers), which follow the same repeating building block: each convolutional layer is followed by batch normalisation, a ReLU activation, and a $2 \times 2$ max-pooling operation. The 3-layer model includes only the first three convolutional blocks (ending at 128 channels), while deeper models extend this pattern by appending additional convolutional blocks. Specifically, the 7-layer model adds two 1024-channel convolutional layers beyond the 5-layer configuration (one with pooling, one without), the 10-layer model includes five such 1024-channel blocks (two with pooling, three without), and the 13-layer model adds eight 1024-channel blocks in total (the first two with pooling, the rest without). All models conclude with a fully connected head comprising a linear layer with 128 units, ReLU activation, dropout ($p = 0.5$), and a final linear classification layer outputting two logits.

All CNN models were trained for 40 epochs using the Adam optimiser (learning rate $1 \times 10^{-3}$) with cross-entropy loss. After each epoch, model performance was evaluated on the test subset. To ensure fair comparison, accuracy-related metrics were computed only on un-augmented data; the augmented data were used solely during training and never in the evaluation phase. The best model epoch was chosen as the epoch with the maximum F1 score.

For benchmarking, Audioset 0, AST Set, was used to train a fine-tune-based AST model with a two-layer classifier head on its 3000 training segments and to evaluate on the 500-segment test split, to provide a direct comparison with the prior art [15] of this study.

The five depths (L3, L5, L7, L10, L13) were chosen as an approximately log-spaced depth sweep designed to

**Table 1** Augmentation operators and parameter settings. Random values are drawn independently for each 5-s clip within the indicated range

| Augmentation | Practical effect on training data | Hyperparameter ($\alpha$) | Drawn range | Audioset ID |
|---|---|---|---|---|
| Pitch-shift | Simulates natural fundamental-frequency variation or slight recorder speed error while preserving timing | 0.1 semitones | $\{-0.1, +0.1\}$ | 3,6 |
| Time-stretch | Mimics faster or slower vocal delivery (tempo changes) and clock drift, without changing pitch | 0.1 | 0.9–1.1 × | |
| Gaussian noise | Adds low-level broadband noise | 0.005 | $N(0, 0.005^2)$ | |
| Time-shift | Moves the call forward or backward inside the 5-s window | 0.2 (fraction) | ± 20% clip length | 2,3,5,6 |
| Volume scale | Emulates different source-to-mic distances and recorder gain settings | 0.2 | 0.8–1.2 × | |

**Table 2** Composition of the audio datasets

| ID | Audioset name | No. of segments | Possum call ratio (%) | Non-possum audio segments source | Train/test | Augmentation techniques |
|---|---|---|---|---|---|---|
| 0 | AST Set | 3500 | 37 | Field record (100%) | 85/15 | No |
| 1 | Untargeted 1X | 2935 | 42 | | 75/25 | No |
| 2 | Untargeted 2X | 5870 | 42 | | | TSFT, VS |
| 3 | Untargeted 3X | 8805 | 42 | | | TSFT, VS, PS, TS, GN |
| 4 | CMCM 1X | 2935 | 42 | Field record (75%)<br>Hard negative (25%) | | No |
| 5 | CMCM 2X | 5870 | 42 | | | TSFT, VS |
| 6 | CMCM 3X | 8805 | 42 | | | TSFT, VS, PS, TS, GN |

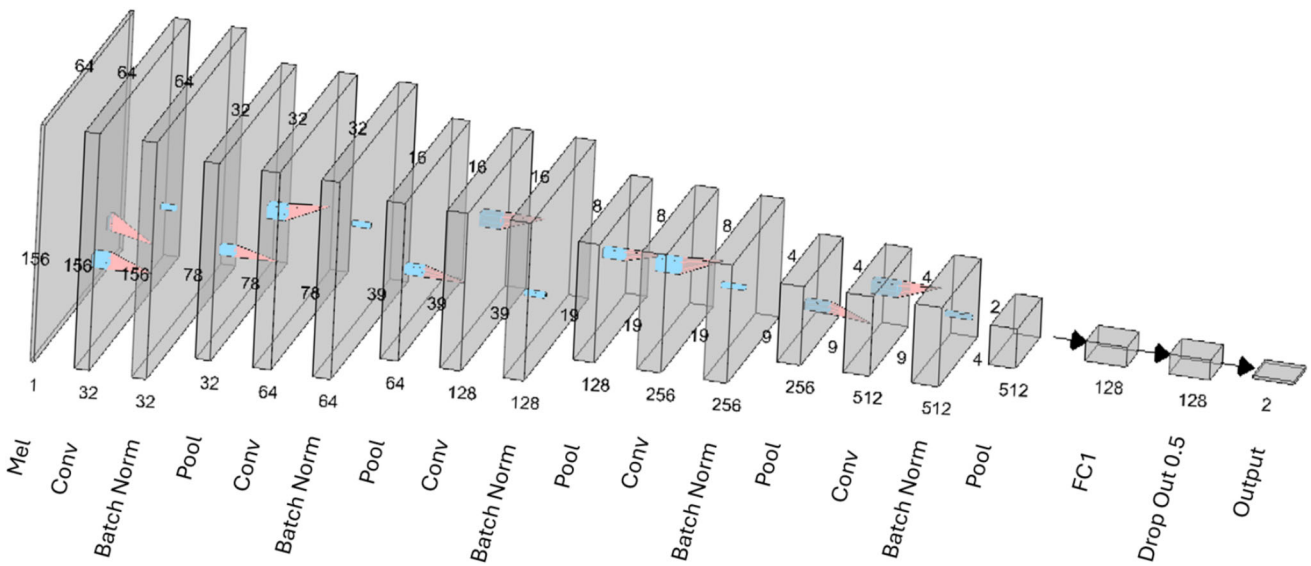


**Fig. 2** Architecture of the 5-layer CNN, representative of all models used. Other variants (3, 7, 10, 13 layers) follow the same block structure with different numbers of convolutional layers

characterise the accuracy–capacity trade-off, broadly bracketing the VGG family of 11–19 weight-layer networks [28] commonly used as audio-spectrogram CNN baselines. The Quadro P2000 (5 GB VRAM) accommodated all five architectures at the batch size used, so the upper depth was not memory-constrained. A more exhaustive hyperparameter search, for example, Bayesian optimisation jointly over depth, channel width, kernel size, and dropout rate, is a natural follow-up and may yield further gains, but is left to future work because the present study is focused on the effect of the CMCM intervention itself rather than on architectural optimisation.

## 5 Results and Discussion

Table 3 shows the list of bird species identified by the probe model when presented with possum calls. Out of 1237 possum calls, 222 were misclassified as Barn Owl, 127 as Bicolored Wren and 108 as Yellow-bellied Sapsucker. Of these three, only the Barn Owl is found in New Zealand. Little Owl, though not as highly confused with possum as the other bird species, is also of interest since it is also found in New Zealand and vocalises at night. Calls of these four birds were obtained from Xeno-Canto, manually verified and included as hard negatives in the CMCM datasets as described in Sect. 3.

Figure 3 illustrates the mel spectrograms of possum chitter and screech compared with those of the four bird species. The spectrograms of possum chitter match closely with those of the yellow-bellied sapsucker, little owl and bicolored wren. Likewise, possum screech is similar to barn owl and little owl screeches. It is interesting to note that these calls are quite distinct to our hearing. However, since a mel spectrogram does not contain phase information, these sounds are less distinguishable from each other. This spectral similarity confirms why these species are the primary sources of false positives, and validates the CMCM approach of explicitly mining them as hard negatives.

Figure 4 illustrates the relationship between the size of the training dataset and the mean accuracy and F1 score for both untargeted and CMCM data, averaged across all five CNN architectures. The error bars cover one standard deviation on either side of the mean. Validation was performed exclusively on un-augmented audio segments to evaluate the models' performance on original sounds.

The trends for both accuracy and F1 score on the validation subsets generally improve with the size of the training data. An exception is the mean F1 score for Untargeted audiosets, which declined slightly when the number of training segments increased from ~ 4,400 (2 × augmentation) to ~ 6,600 (3 × augmentation). This suggests that further augmentation did not yield additional performance gains for F1 in this case. The maximum F1 score (99.03%) and maximum accuracy (98.1%) were achieved on the Untargeted 2X and CMCM 2X audiosets, respectively.

The mean F1 score percentage for the fivefold cross-validation is shown in Fig. 5. Each data point represents the average score from five cross-validation runs. The highest mean F1 scores are achieved by the models trained on the CMCM 3X audioset, followed by those trained on CMCM 2X. Overall, models trained with hard negatives using the CMCM method yielded higher mean F1 scores compared to those trained with untargeted non-possum audio. These results suggest that exposure to more challenging negative samples during training improves a model's ability to distinguish possum calls, resulting in greater robustness in both accuracy and precision. The L3 CNN model produced the lowest F1 score. The highest score is achieved by the L10 CNN.

The hard-negative clips used in training are drawn from Xeno-Canto, where each recording is contributed as a single-species clip rather than a field soundscape, and the possum-positive clips are from the public Kaggle dataset of McEwen et al.; both repositories are publicly accessible and the full hard-negative set is reproducible by any reader following the procedure in Sect. 3. Because brushtail possums and the two NZ-occurring CMCM owls (Barn Owl, Little Owl) are nocturnal, within-clip overlap is in principle possible in field soundscapes, but by design of the source repositories it is essentially absent from the training data, so the model is never shown a clip in which an owl is present and the clip is labelled non-possum. In field soundscapes the situation is asymmetric: owl calls are often delivered in extended sequences while possum vocalisations are episodic, so most owl-bearing windows contain no possum and are correctly rejected, while windows that do contain a possum call retain its distinctive signature. The spectrograms in Fig. 3 show that the possum and owl signals are similar but not identical, and the high cross-validation F1 scores in Fig. 5 confirm that the model discriminates between them on held-out data. Frame-level analysis of within-clip co-occurrence, including the native morepork (Ninox novaeseelandiae), which was not included in the CMCM set, is left as future work.

The trained models were then evaluated on the "Real-World" dataset described in Sect. 4.1 to determine whether possum calls could be correctly identified without generating excessive false alarms. The results are summarised in Table 4. It is important to note that this evaluation already constitutes a test against largely unseen confusers. Of the four bird species used to construct the CMCM training data (Barn Owl, Bicolored Wren, Yellow-bellied Sapsucker, Little Owl; see Table 3), only Barn Owl and Little Owl occur in New Zealand at all, so the great majority of bird vocalisations present in the one-hour NZ native-forest recording necessarily come

**Table 3** Hard negatives mined through CMCM

| Birds vocalising possum Hard-negative calls | Number of confusions | Average confidence | Found in New Zealand | Vocalise at night |
|---|---|---|---|---|
| Barn Owl | 222 | 0.42 | Yes | Yes |
| Bicolored Wren | 129 | 0.25 | No | No |
| Yellow-bellied Sapsucker | 108 | 0.21 | No | No |
| Fasciated Wren | 80 | 0.23 | No | No |
| Mistle Thrush | 56 | 0.22 | No | No |
| Cactus Wren | 51 | 0.22 | No | No |
| Yellow-headed Blackbird | 42 | 0.29 | No | No |
| Northern Hawk Owl | 24 | 0.28 | No | No |
| Cardinal Woodpecker | 23 | 0.2 | No | No |
| Eurasian Jay | 20 | 0.26 | No | No |
| Band-backed Wren | 19 | 0.27 | No | No |
| Red-throated Ant-Tanager | 19 | 0.2 | No | No |
| Nanday Parakeet | 14 | 0.25 | No | No |
| Canada Jay | 10 | 0.18 | No | No |
| Black Woodpecker | 9 | 0.25 | No | No |
| Grey Treepie | 8 | 0.2 | No | No |
| Western Capercaillie | 6 | 0.17 | No | No |
| Little Owl | 6 | 0.24 | Yes | Yes |
| Eurasian Bullfinch | 6 | 0.3 | No | No |
| White-bellied Treepie | 6 | 0.21 | No | No |
| Arrow-marked Babbler | 6 | 0.21 | No | No |

The number of confusions reflects the number of possum call segments that were confused with each bird species. The total number of possum call segments was 1237

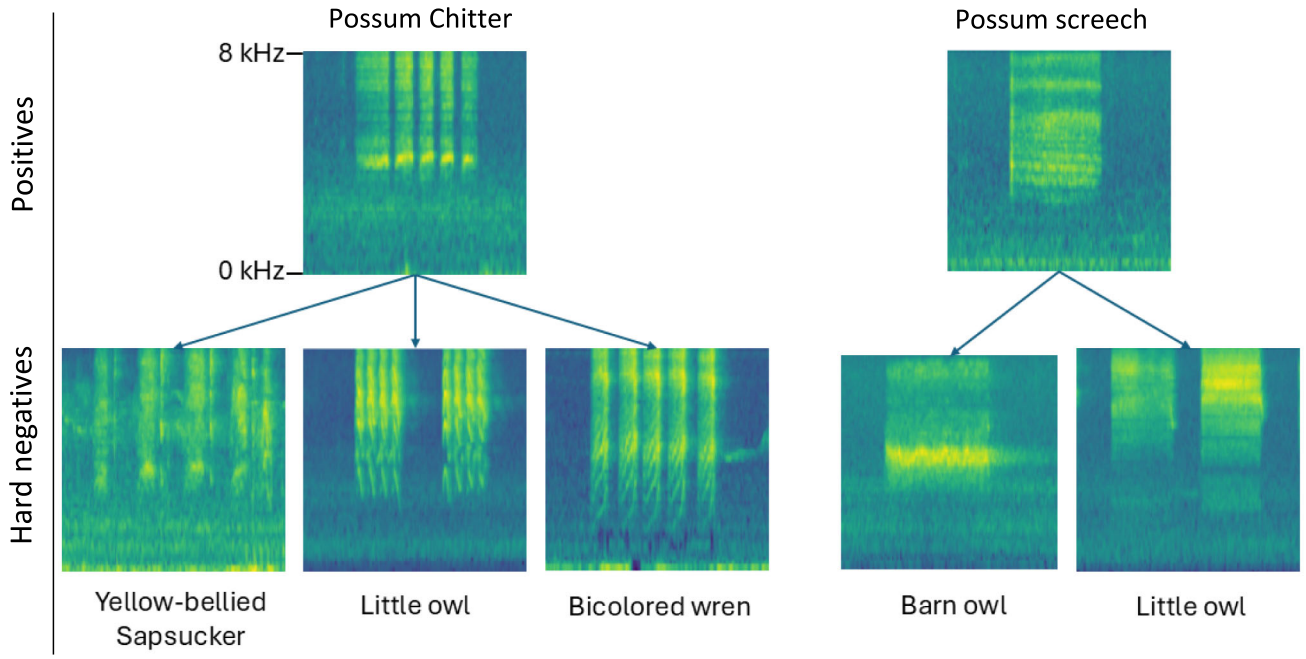


**Fig. 3** Samples of the mined hard negatives through CMCM in the mel spectrogram. Each mel spectrogram is approximately 3 s

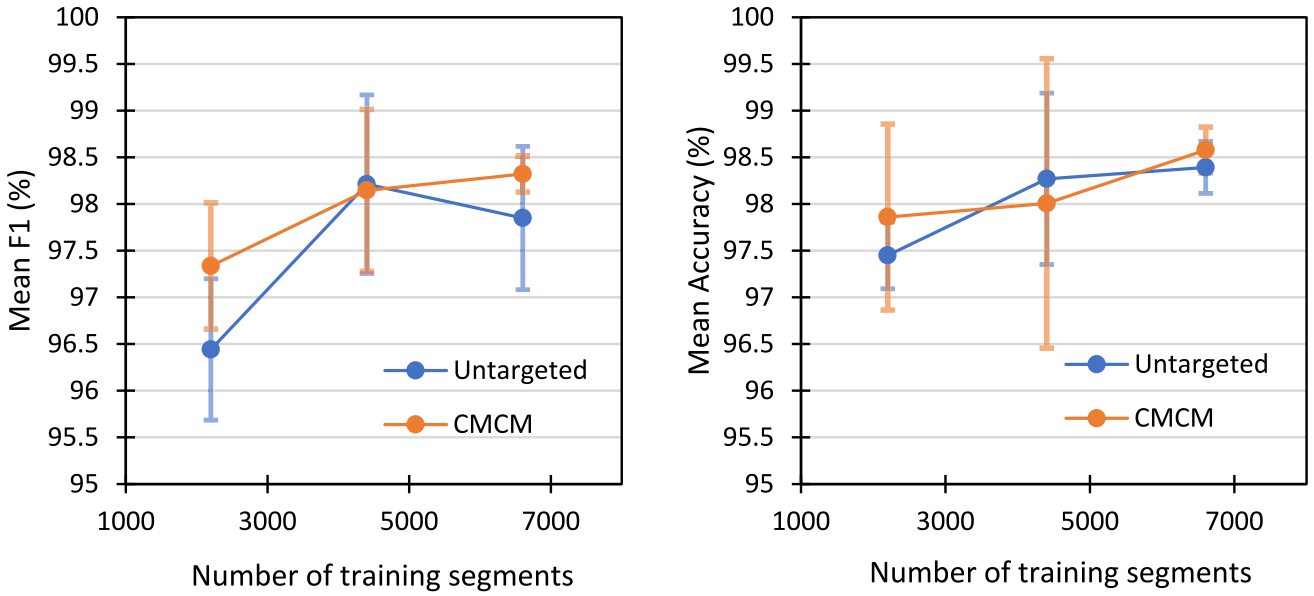


**Fig. 4** Model performance evaluation in terms of F1 (%) (**A**) and Mean Accuracy (%) (**B**). The results are based on un-augmented validation subsets. Each data point reflects the mean result of all models with CNN layers 3, 5, 7, 10, and 13

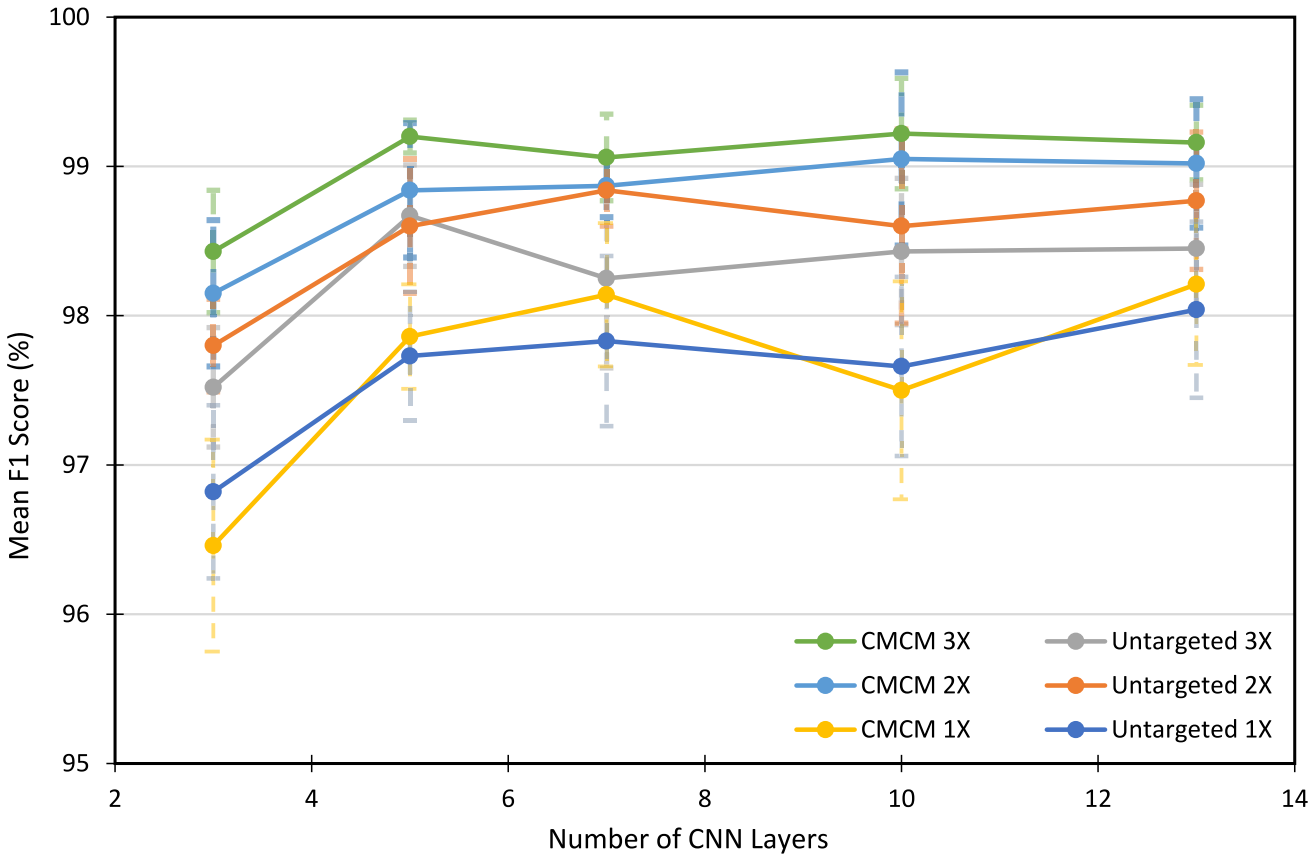


**Fig. 5** FiveFold stratified cross-validation results. The F1 Scores (%) are based on un-augmented data in the validation fold

**Table 4** Real-world simulation dataset results

| Audioset | Number of CNN layers | Real-world audio record (number of possum call detections) | | |
|---|---|---|---|---|
| | | NZ forest record with 0 possum call | Possum call record one with 9 true possum calls | Possum call record two with 7 true possum calls |
| CMCM 3X | 10 | ✓ (0) | ✓ (4) | ✓ (1) |
| CMCM 3X | 7 | ✓ (0) | ✗ (0) | ✓ (3) |
| CMCM 3X | 5 | ✓ (0) | ✗ (0) | ✓ (2) |
| CMCM 3X | 13 | ✗ (7) | ✓ (5) | ✓ (2) |
| CMCM 2X | 10 | ✓ (0) | ✓ (1) | ✓ (2) |
| CMCM 2X | 7 | ✓ (0) | ✗ (0) | ✓ (2) |
| CMCM 2X | 5 | ✓ (0) | ✗ (0) | ✓ (1) |
| CMCM 2X | 13 | ✗ (1) | ✗ (0) | ✓ (3) |
| CMCM 2X | 3 | ✓ (0) | ✗ (0) | ✗ (0) |
| CMCM 1X | 7 | ✗ (5) | ✓ (8) | ✓ (3) |
| CMCM 1X | 5 | ✗ (19) | ✓ (3) | ✓ (4) |
| CMCM 1X | 3 | ✓ (0) | ✗ (0) | ✓ (1) |
| CMCM 1X | 13 | ✓ (0) | ✗ (0) | ✓ (3) |
| CMCM 1X | 10 | ✗ (12) | ✗ (0) | ✓ (5) |
| Untargeted 3X | 7 | ✗ (418) | ✗ (13) | ✗ (9) |
| AST model | – | ✗ (27) | ✗ (13) | ✓ (1) |
| Untargeted 3X | 5 | ✗ (280) | ✗ (13) | ✗ (9) |
| Untargeted 3X | 3 | ✗ (811) | ✗ (13) | ✓ (25) |
| Untargeted 3X | 13 | ✗ (596) | ✗ (13) | ✗ (12) |
| Untargeted 3X | 10 | ✗ (301) | ✗ (13) | ✓ (7) |
| Untargeted 2X | 7 | ✗ (732) | ✗ (13) | ✗ (14) |
| Untargeted 2X | 5 | ✗ (755) | ✗ (13) | ✗ (23) |
| Untargeted 2X | 3 | ✗ (901) | ✗ (13) | ✗ (34) |
| Untargeted 2X | 13 | ✗ (725) | ✗ (13) | ✗ (12) |
| Untargeted 2X | 10 | ✗ (660) | ✗ (13) | ✗ (20) |
| Untargeted 1X | 10 | ✗ (862) | ✗ (13) | ✗ (20) |
| Untargeted 1X | 13 | ✗ (957) | ✗ (13) | ✗ (25) |
| Untargeted 1X | 7 | ✗ (737) | ✗ (13) | ✗ (25) |
| Untargeted 1X | 5 | ✗ (910) | ✗ (13) | ✗ (27) |
| Untargeted 1X | 3 | ✗ (715) | ✗ (12) | ✗ (21) |

Tick mark (✓) means no false alarms; cross mark (✗) means at least one false alarm or failure to detect any possum call. The numbers in the brackets reflect the number of possum call detection alarms in each audio file

from species that were not in the training set. The zero-false-positive performance of the L10 CMCM 3X model on this hour of audio therefore reflects generalisation to unseen confusers rather than retrieval of the trained classes. Almost all models trained using CMCM outperformed those trained on untargeted audiosets. The L10 CNN models trained on CMCM 2X and CMCM 3X datasets excelled at not producing false alarms. In detecting possum calls, the model trained on CMCM 3X detected two more calls than the model trained on CMCM 2X; these results are highlighted in the table. In contrast, the same models trained with untargeted datasets generated hundreds of false alarms when presented with the hour-long forest audio containing no possum calls. Interestingly, these performance differences are not reflected in the F1 scores obtained when the models are evaluated on the audiosets used for training. This observation emphasises the importance of evaluating machine learning models on data drawn from sources different from those used during training.

The superiority of the L10 network trained with CMCM 3X is more clearly demonstrated through the following performance metric:

$$\text{PenaltyScore} = \frac{\text{TP}}{1 + \text{FP}} \tag{1}$$

where TP and FP are the number of True Positives and False Positives, respectively. This penalty score penalises false positives and rewards true positives.

Figure 6 ranks every network–audioset combination in descending order of this score. L10 trained on the CMCM 3X audioset achieved the highest penalty score of 5, with no false positives while detecting five of the sixteen true possum calls. This is followed by other models trained on CMCM audiosets, all of which had zero false alarms and therefore retained the benefit of their true-positive counts. In contrast, models trained on untargeted audiosets appear near the bottom of the chart due to large numbers of false detections. It should be noted that the CMCM-trained models detected fewer total possum calls than some untargeted models, reflecting a conservative bias that prioritises precision over recall, a desirable trade-off in field monitoring where false alarms waste limited operational resources.

The baseline AST model produced approximately 30 false alarms when evaluating the Real-World audiosets and detected fewer possum calls compared to models trained on audiosets with hard negatives using the CMCM method. It has a penalty score of 0.31.

The computational costs of each model, in terms of the mean number of floating-point operations (MFLOPs) required for a single 5-s inference and the corresponding latency, are shown in Fig. 7. These values are obtained on an NVIDIA Quadro P2000 GPU. As expected, the transformer-based AST model demands the most computational resources at over 1.3 GFLOPs as well as the slowest in inferencing at roughly 400 ms per inference. In comparison, the L3 CNN demanded fewer than 100 MFLOPs and completes inference in under 10 ms, whereas the L13 requires ~ 400 MFLOPs and ~ 35 ms. These results indicate that the AST model may exceed practical limits for edge deployment unless specialised acceleration is available.

### 5.1 Generalisation to Unseen Confusers

A natural concern with any hard-negative mining method is whether the resulting classifier has merely learned to reject the specific negative classes seen during training, or whether it generalises to the broader population of acoustically similar non-target sounds. The real-world evaluation in Table 4 already addresses this concern, but the design choice that makes it an unseen-confuser test was not made explicit. The four bird species used to populate the CMCM datasets (Sect. 4.1) were selected on the basis of how frequently BirdNET confused them with possum calls, not on the basis of their presence in New Zealand. Two of the four (Bicolored Wren, Yellow-bellied Sapsucker) do not occur in New Zealand at all, and the one-hour DOC native-forest recording therefore consists almost entirely of NZ avifauna that the model had not seen during training.

The L10 CMCM 3X model produced zero false alarms across this hour of acoustically rich, geographically disjoint audio. By construction, the great majority of bird detections (or non-detections) on the NZ soundscape are responses to confusers the model had never seen. This indicates that the CMCM-trained model has not memorised the specific spectro-temporal signatures of the four training confusers, but has instead learned a more general decision boundary around the possum call manifold. By contrast, the AST baseline**,** which was trained without CMCM-mined hard negatives**,** produced approximately 30 false alarms on the same recording, indicating that exposure to confusable bird vocalisations during training, rather than model capacity, is the determining factor for false-positive suppression.

### 5.2 Edge-Deployment Considerations

The computational profile reported in Fig. 7 already addresses the compute and latency aspects of edge deployment. Because the recommended L10 architecture has strictly fewer layers than L13 with the same channel widths in the deeper blocks, its inference cost is upper bounded by the L13 figures of approximately 400 MFLOPs and 35 ms per five-second window on the Quadro P2000, less than one-third of the AST baseline's FLOP cost and more than an order of magnitude lower in latency. Mel-spectrogram preprocessing (1024-point short-time Fourier transform (STFT) with 512-sample hop, 64 mel bins) adds approximately 2 MFLOPs per clip, well under a millisecond on any ARM-class CPU and is not a deployment bottleneck. The CNN parameter count is dominated by the deeper 1024-channel blocks and is comparable to, though below, the AST baseline in floating-point form; the decisive advantage of the proposed approach is therefore computational efficiency rather than raw storage. Standard post-training INT8 quantisation offers an additional 4 × reduction in model size with negligible accuracy loss for CNNs of this scale and is a natural step before field deployment. On-device latency, memory, power, and energy measurements on a specific platform are identified as follow-up work in Sect. 6.

## 6 Conclusions and Future Work

In this paper, the Cross-Model Confusion Mapping (CMCM) was introduced to improve the training of Convolutional

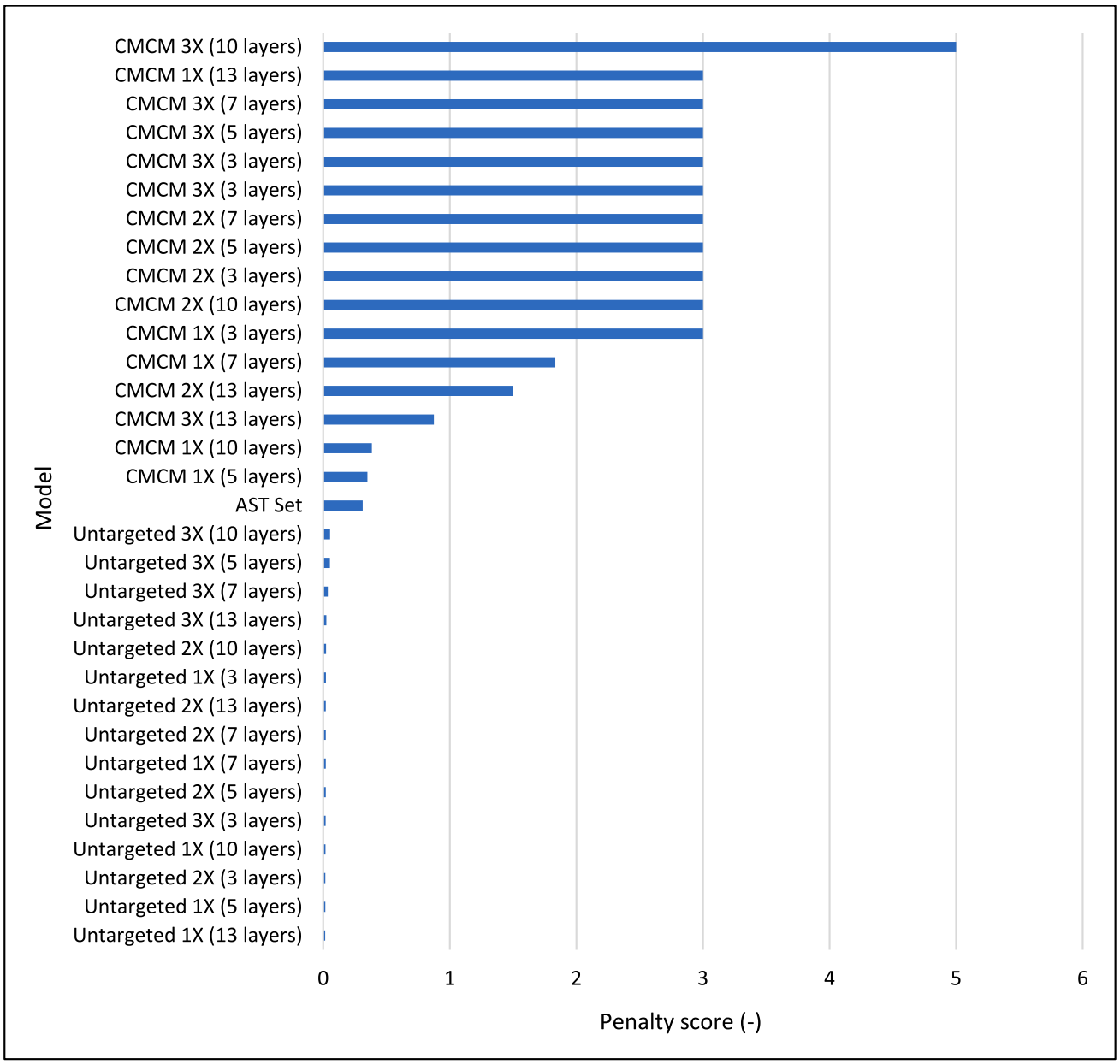


**Fig. 6** Penalty score for all the models

Neural Networks to detect possum calls with a very low false-positive rate. Experimental results with 3, 5, 7, 10, and 13-layer CNNs showed that the 10-layer CNN trained on the CMCM 3X augmented audioset produced the best overall results. Furthermore, not a single false positive resulted when it was presented with a one-hour forest recording. It also correctly detected five of sixteen possum calls in two unseen possum-positive clips, yielding the highest precision. Inference profiling showed that this model completed a forward pass two orders of magnitude faster than the AST benchmark while requiring less than one-third of its computational cost in terms of FLOPs. These findings demonstrate that our targeted hard-negative mining CMCM method, when combined with moderate data augmentation, can deliver a detector that is both accurate and computationally efficient. This makes it suitable for deployment on low-power edge computing devices.

A limitation of this study is that the real-world evaluation relied on a small number of publicly available recordings; larger-scale field trials are needed to confirm these findings across diverse habitats and seasons. Since the current models occasionally missed previously unheard possum calls,

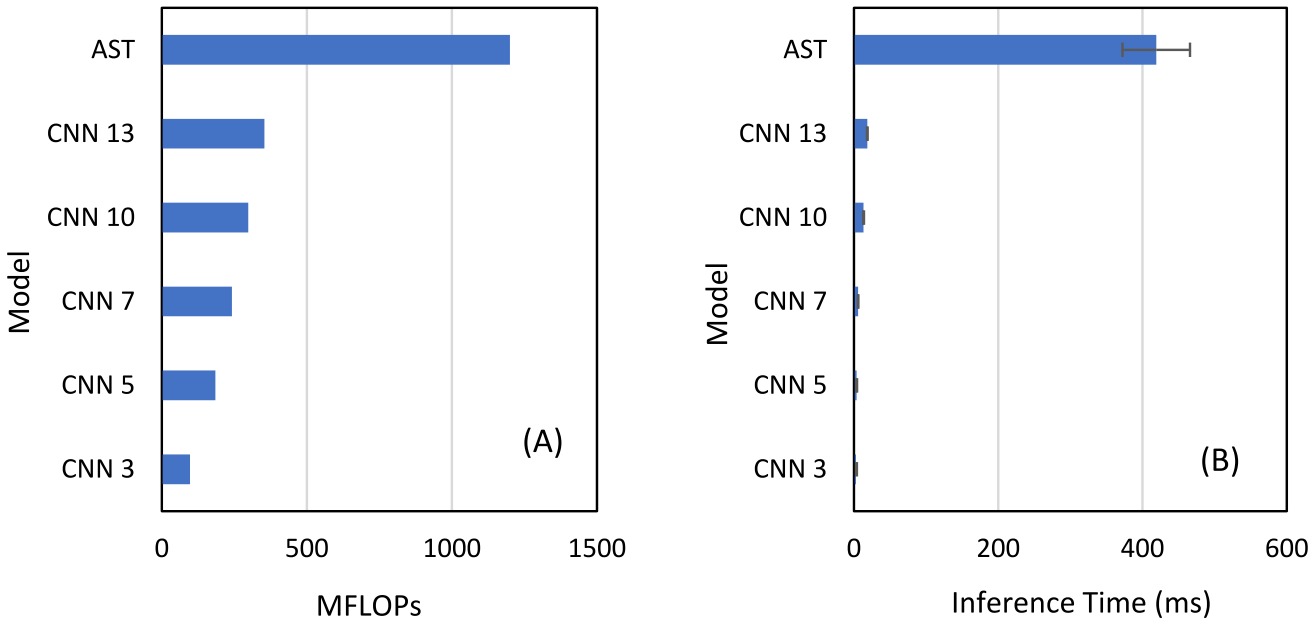


**Fig. 7** Computational footprint of the six possum-detection networks: **A** Millions of floating-point operations (MFLOPs) per inference; **B** Inference latency (ms) on an NVIDIA Quadro P2000 GPU

expanding the library of wild (non-captive) possum vocalisations is important. Other directions for future work include utilising the CMCM method for other pest mammals such as stoats and rats. On-device latency, peak memory, power draw, and energy per detection on representative low-power platforms (for example Raspberry Pi 4, Jetson Nano, or Coral Dev Board), together with the impact of INT8 post-training quantisation, are immediate follow-up work. Joint hyperparameter optimisation of depth, channel widths, kernel sizes and dropout, beyond the discrete depth sweep used here, is another natural extension that may further improve the precision/latency trade-off.

**Acknowledgements** The authors acknowledge Auckland University of Technology for supporting this research as part of a postdoctoral programme.

**Author Contributions** A.G. contributed to the conceptualisation, provided supervision and validation, and reviewed the manuscript. M.A.B. conceived the study, developed the CMCM methodology, wrote the software, performed all formal analysis and data curation, wrote the main manuscript text, and prepared all figures and tables. E.L. provided supervision and funding acquisition, and reviewed the manuscript. All authors read and approved the final manuscript.

**Funding** Open Access funding enabled and organized by CAUL and its Member Institutions.

**Data Availability** The datasets generated and/or analysed during the current study are available from the corresponding author on reasonable request.

## Declarations

**Conflict of interest** The authors declare no competing interests.



## References

1. Allen, R.B., Lee, W.G.: Biological invasions in New Zealand. Springer, Berlin (2006)
2. King, C., Forsyth, D.: The handbook of New Zealand mammals. Csiro Publishing (2021)
3. Latham, A.D.M., Latham, M.C., Norbury, G.L., Forsyth, D.M., Warburton, B.: A review of the damage caused by invasive wild mammalian herbivores to primary production in New Zealand. N. Z. J. Zool. **47**, 20–52 (2020). https://doi.org/10.1080/03014223.2019.1689147
4. Livingstone, P., Hancox, N., Nugent, G., de Lisle, G.: Toward eradication: the effect of Mycobacterium bovis infection in wildlife on the evolution and future direction of bovine tuberculosis management in New Zealand. N. Z. Vet. J. **63**, 4–18 (2015). https://doi.org/10.1080/00480169.2014.971082

5. Jones, C., Barron, M., Warburton, B., Coleman, M., Lyver, P.O., Nugent, G.: Serving two masters: reconciling economic and biodiversity outcomes of brushtail possum (*Trichosurus vulpecula*) fur harvest in an indigenous New Zealand forest. Biol. Conserv. **153**, 143–152 (2012). https://doi.org/10.1016/j.biocon.2012.04.016
6. Russell, J.C., Innes, J.G., Brown, P.H., Byrom, A.E.: Predator-free New Zealand: conservation country. Bioscience **65**, 520–525 (2015)
7. Ministry for Primary Industries: Introduction to animal pest monitoring: a guide to monitoring animal pests in New Zealand. Department of Conservation and Ministry for Primary Industries, Wellington (2020)
8. Sweetapple P, Nugent G (2011) Chew-track-cards: a multiple-species small mammal detection device. New Zealand J. Ecol., 153–162
9. Harley, D., Eyre, A.: Ten years of camera trapping for a cryptic and threatened arboreal mammal–a review of applications and limitations. Wildlife Res. **51**, 23054 (2024)
10. Harley DK, Holland GJ, Hradsky BAK, Antrobus JS (2014) The use of camera traps to detect arboreal mammals: lessons from targeted surveys for the cryptic Leadbeater's Possum Gymnobelideus leadbeateri. Camera trapping: Wildlife management and research 233–243
11. Potter, L.C., Brady, C.J., Murphy, B.P.: Accuracy of identifications of mammal species from camera trap images: a northern Australian case study. Austral Ecol. **44**, 473–483 (2019). https://doi.org/10.1111/aec.12681
12. Gong Y, Chung Y-A, Glass J (2021) AST: Audio Spectrogram Transformer
13. Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M, Adam H (2017) MobileNets: efficient convolutional neural networks for mobile vision applications
14. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp 770–778
15. McEwen, B., Bainbridge-Smith, A., Gutschmidt, S., Green, R., Atlas, J.: An invasive species model and dataset for bioacoustic monitoring of common brushtail possum. N. Z. J. Ecol. **48**, 1–6 (2024)
16. Merchan, F., Guerra, A., Poveda, H., Guzmán, H.M., Sanchez-Galan, J.E.: Bioacoustic classification of antillean manatee vocalization spectrograms using deep convolutional neural networks. Appl. Sci. **10**, 3286 (2020). https://doi.org/10.3390/app10093286
17. LeBien, J., Zhong, M., Campos-Cerqueira, M., Velev, J.P., Dodhia, R., Ferres, J.L., Aide, T.M.: A pipeline for identification of bird and frog species in tropical soundscape recordings using a convolutional neural network. Ecol. Inform. **59**, 101113 (2020). https://doi.org/10.1016/j.ecoinf.2020.101113
18. Allen, A.N., Harvey, M., Harrell, L., Jansen, A., Merkens, K.P., Wall, C.C., Cattiau, J., Oleson, E.M.: A convolutional neural network for automated detection of humpback whale song in a diverse, long-term passive acoustic dataset. Front. Mar. Sci. (2021). https://doi.org/10.3389/fmars.2021.607321
19. Bucher, M., Herbin, S., Jurie, F.: Hard negative mining for metric learning based zero-shot classification. In: Hua, G., Jégou, H. (eds.) Computer Vision – ECCV 2016 Workshops, pp. 524–531. Springer International Publishing, Cham (2016)
20. Kahl, S., Wood, C.M., Eibl, M., Klinck, H.: BirdNET: a deep learning solution for avian diversity monitoring. Eco. Inform. **61**, 101236 (2021)
21. Vellinga W-P, Planqué R (2015) The Xeno-canto Collection and its Relation to Sound Recognition and Classification. CLEF (Working Notes) 1391:
22. McEwen B, Bainbridge-Smith A, Gutschmidt S, Green R, Atlas J (2024) Invasive Species NZ, Kaggle
23. Barzegar MA (2025) https://github.com/mabiran/Audio-Annotator
24. (2022) 1 Hour New Zealand Forest at Night Ambience | Study, work, relax, immerse yourself in nature
25. (2016) Brushtail possum making wierd noises. https://www.youtube.com/watch?v=nqZtuw-O-G8. Accessed 13 June 2025
26. (2020) Common Brushtail Possum Sounds - Scary growling calls and scurrying noises. https://www.youtube.com/watch?v=IouFOQDr4Sc. Accessed 13 June 2025
27. Paszke A (2019) Pytorch: An imperative style, high-performance deep learning library. arXiv preprint arXiv:191201703
28. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition