Western Jackdaw Call Classification in Noisy Environments Using CNNs

Anonymous submission to Interspeech 2024

Abstract

The western jackdaw (Corvus/Coloeus monedula) is a passerine bird found across Europe. Automated detection and classification of jackdaw calls from audio recordings can support population monitoring and behavioral studies. However, background noise presents significant challenges. This research presents multiple deep-learning approaches for jackdaw call classification robust to realistic environmental noise. Experiments are performed with multiple deep-learning models using different features including custom convolutional neural network (CNN) models using MFCC and spectrogram features, pre-trained BirdNet, InceptionV3, Xception, ResNet50 models using spectrograms, LSTM-based RNN models using MFCCs and pre-trained transformer models using raw waveforms(Wav2Vec2). Tests performed on a manually curated dataset of jackdaw calls and noise segments extracted from field recordings achieve the best performance of 98% accuracy on the validation set using custom CNN and BirdNet models. Further manual validation of extracted Jackdaw calls on unlabeled raw field data shows a precision score of 93%. This research also presents data balancing and aggressive noise filtering to improve model generalization under varying real-world noise.

Index Terms: bird call recognition, MFCCs, spectrogram analysis, deep learning

1. Introduction

Automated bird species recognition from recordings is critical for scalable avian ecology research and conservation efforts [1, 2]. Manual field surveys are limited in scope, time-intensive, and susceptible to observer biases. In contrast, automated methods based on deep neural networks now rival expert-level performance in biodiversity monitoring from audio data [3]. Specifically, convolutional neural networks (CNNs) demonstrate state-of-the-art capabilities in classifying bird vocalizations to species [4, 5, 6]. Key enablers include using spectrogram image representations of audio data as input [7], thereby leveraging transfer learning from extensive image recognition research[8, 9].

However, background noise remains the primary impediment to accurate classification, causing pervasive false detections and inaccuracies [10]. Interfering sounds like wind, rain, machinery, human activity, and calls from other species present in real-world field recordings introduce major challenges. Addressing this requires developing robust algorithms specifically tailored for noisy conditions frequently encountered in ecosystem monitoring. While prior innovations demonstrate promise in improving noise resilience on controlled single-species audio datasets, they struggle to bridge the gap to uncontrolled natural soundscapes with complex noise profiles [6, 11].

Current top-performing approaches draw from a fusion of techniques to mitigate background noise issues. Adaptive segmentation algorithms show capabilities to extract target species vocalizations from long recordings [6]. Data augmentation through pitch shifting, time warping, mixing samples, adds crucial variation to limited training datasets[12, 11]. Time-frequency transformations like mel-spectrograms and wavelet transform better match bird acoustic qualities [13]. Ensemble approaches using multiple parallel CNN architectures build consensus and reduce outlier predictions [14]. Adding recurrent neural network and attention layers also assists by modeling temporal context[15, 16]. Noise reduction as a preprocessing step can further suppress interference before classification [17, 14].

However, gains achieved from these methods remain constrained. Performance degradation persists in practice due to mismatches between training and deployment conditions. Additional research through the lens of multidisciplinary collaboration combining ecology, acoustics, and machine learning is essential to realize robust species classification amidst real-world noisy soundscapes.

Corvids (Family: Corvidae) contain the crows, rooks, jackdaws, ravens, jays, magpies, treepies and nutcrackers. They are known for their complex vocalizations and communication skills. Audio recognition technology, where it can be readily used to identify and analyze corvid vocalizations, may represent a cornerstone to unlocking further research into their ecology. Our research contributes towards this aim, developing a tailored neural network pipeline for recognizing western jackdaw (Corvus/Coloeus monedula) vocalizations under varying noise. We integrate state-of-the-art techniques from literature to design an accurate jackdaw classifier serving as a case study for conservation-focused bioacoustic monitoring. The novelty of this research apart from building the first ever dedicated robust jackdaw call recognition model comparing different state-of-the-art deep learning architectures is to test it on real-world noisy field recordings.

2. Methodology

The most popular audio features for any audio recognition task are the Mel-frequency cepstral coefficients (MFCCs). This feature extraction technique applies the mel-scale filter-banks to amplify lower frequency components on the logarithmic power spectrum and then transforms these to cepstral coefficients using the inverse Fourier transform.

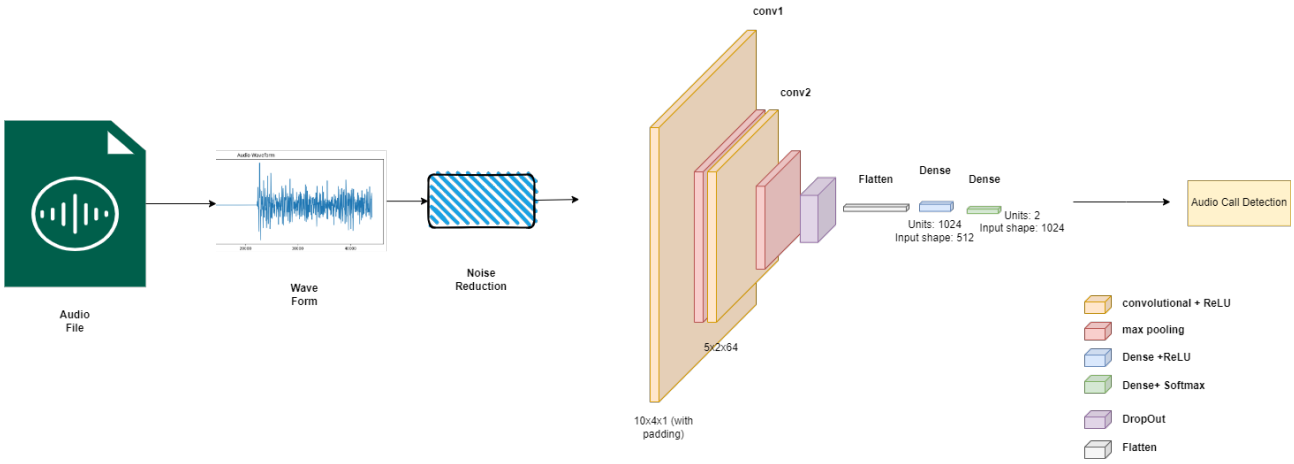

Figure 1: *Custom CNN Model Architecture with audio features*

2.1. Noise Reduction

The noise reduction technique employed in this research utilizes a low-pass filtering approach to enhance the quality of audio signals. Initially, the Fast Fourier Transform (FFT) is applied to convert the input audio signal from the time domain to the frequency domain. Subsequently, a copy of the FFT of the signal is created, and a threshold-based filtering operation is performed, where frequency components beyond a specified threshold are set to zero. This threshold is checked for values ranging from 0.1 to 1, and is set to 0.3 based on experiments showing optimal performance. This effectively eliminates high-frequency noise from the signal. Following this, the inverse FFT is applied to transform the filtered signal back to the time domain, yielding a cleaner version of the original audio signal. The benefits of this technique are manifold: it enhances signal quality by reducing distortion, improves feature extraction accuracy, particularly when using methods like Mel-frequency cepstral coefficients (MFCCs), and leads to better performance of subsequent processing tasks such as training Convolutional Neural Networks (CNNs). Additionally, the noise reduction enhances the robustness of models to environmental variations, contributing to better generalization on unseen data. Overall, this technique serves as a valuable preprocessing step in audio signal processing workflows, facilitating more accurate and effective analysis and modeling of audio data. Given:

- Input audio signal: y
- Sample rate: sr
- Threshold for filtering: th

1. **Compute the Fast Fourier Transform (FFT) of the input signal:**

$$y_f = \text{FFT}(y)$$

2. **Define the frequency domain components:**

Total number of samples: $N = \text{int}(sr \times \text{DURATION})$

Frequencies for FFT bins: $xf = \text{fftfreq}(N, \frac{1}{\text{SAMPLE\_RATE}})$

3. **Apply a low-pass filter by setting the frequency components beyond a certain threshold to zero:**

Create a copy of the FFT of the input signal:

$\text{new\_yf} = y_f.\text{copy}()$

Define the middle index: $\text{middle} = \frac{\text{len}(y)}{2}$

Set the frequency components beyond the threshold to zero:

$\text{new\_yf}\,[\text{int}\,(\text{middle} - \text{len}(y) \times th) : \text{int}\,(\text{middle} + \text{len}(y) \times th)] = 0$

4. **Compute the inverse FFT to obtain the filtered signal in the time domain:**

$$\text{new\_y} = \text{IFFT}(\text{new\_yf})$$

$$\text{new\_y} = \text{real}(\text{new\_y})$$

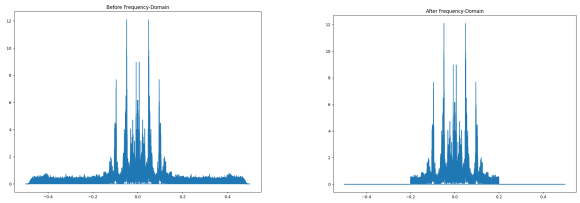

Figure 2: *Before And After Noise Reduction*

2.2. Model Architectures

The custom model used here is a Convolutional Neural Network (CNN) architecture designed for audio classification tasks. It consists of several convolutional layers that apply filters to the input data, extracting meaningful features through a series of convolutions and max pooling operations. The convolutional layers are followed by dropout layers to prevent overfitting during training. The extracted features are flattened and passed through fully connected (dense) layers, culminating in a final layer with two neurons and a softmax activation function. This output layer allows the model to classify the input data into one of the two categories. The model's strength lies in its ability to automatically learn hierarchical representations of the input

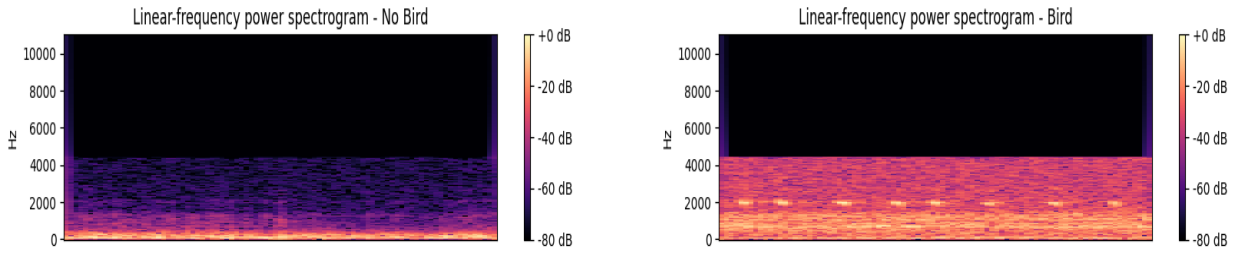

Figure 3: *Linear-frequency Power Spectrogram for No Bird and Jackdaw Call*

data, making it well-suited for tasks like audio classification. CNNs are particularly effective for processing data with spatial or temporal dependencies, such as audio spectrograms or images. By leveraging local connectivity and weight sharing, CNNs can efficiently capture patterns and learn robust feature representations. Additionally, the use of max pooling layers helps to reduce computational complexity and introduces invariance to small shifts or distortions in the input data. Overall, the CNN architecture is a powerful and widely used approach for audio analysis tasks, capable of learning complex patterns and achieving high performance in various audio classification problems.

The BirdNet sound classification model, a pre-trained CNN model for bird audio classification task, employs a dual-spectrogram input representation to provide complementary views of the raw waveform for the neural network. The model accepts audio sampled at 48kHz, and automatically resamples if necessary. Two separate mel-scale spectrograms are extracted using different parameters to capture both low and high-frequency content:

The first spanning 0-3kHz uses a 2048-point FFT with 278 hop size and 96 mel bins. The second from 500Hz-15kHz uses a 1024-point FFT with 280 hop size, also with 96 mel bins. Both have a resulting dimensionality of 96x511. Raw audio is normalized between -1 and 1 before spectrogram conversion. A nonlinear mapping is applied to the magnitude spectra as described in [12] to improve sound event detection. The dual-spectral input provides a rich initial representation of the data to the deep neural network.

The backbone classification architecture is EfficientNetB0-inspired, employing inverted residual blocks with squeeze-and-excitation modules. It processes the multi-resolution input through several convolutional layers, followed by global average pooling to produce a 1024-unit embedding vector representing the input audio clip. This is finally classified using a linear layer to predict bird species. The dual-spectrogram input in conjunction with an efficient deep neural network allows accurate detection and classification of bird vocalizations from raw waveform audio recordings. The model is robust to sampling rate variation and represents both low and high-frequency content useful for identifying bird sounds.

3. Experiments and results

Experiments are performed on the field recordings using multiple deep-learning models.

3.1. Datasets

Models were trained and tested from data collected from a colony of western jackdaws in a rural setting in Sweden. The data were collected using Audiomoth [18] autonomous recording units between March - July 2023, as part of an ongoing bioacoustic study. Recordings were made at 48000Hz with a bit rate of 16 bits in wave format. Recording units were deployed at a range of 5-10m from occupied nest boxes. A part of the dataset ($\tilde{9}$ hours) was manually labeled with calls and other background classes to be modeled as a binary class task. The labeled data consists of 2576 samples (128 mins) of Jackdaw calls and 14867 samples (743 mins) of false positive background audio. This labeled dataset was split in the ratio of 80/20 for the train-test split and a separate cross-validation set within the train set to estimate the best-performing hyperparameters. Further, $\tilde{1}$7.7 hours of raw field recordings are used as unlabelled test data to manually validate the models for future deployment.

3.2. Experimental conditions

Multiple deep-learning models are being implemented for building a Jackdaw call recognition model. This includes both custom models and fine-tuned pre-trained models using both audio features and spectrogram-based image features. The custom CNN models were trained for MFCC audio features as a 1-D CNN model or for spectrogram image features as a 2-D CNN model. Similarly, multiple pre-trained CNN models (like ResNet50, InceptionV3, and Xception), including the popular BirdNet model are fine-tuned using the data. Experiments are also performed on sequence models like LSTMs and pre-trained transformer models.

3.3. Experimental result

As mentioned earlier, the model performances are reported on the held-out test set. The range of hyperparameters tuned for the models is depicted in Table 1. The initial experiments are performed using data balancing and noise reduction techniques as shown in table 2 on the custom CNN model. It is observed that noise reduction is a key step in improving performance and data balancing may not have as much impact as expected. Considering this, multiple models are trained using noise reduction on unbalanced data as shown in table 3. The table shows accuracy (as Acc), precision (as P), recall (as R) and F1-score (as F1) on the labeled test set. The results show that the spectrogram features show slightly better performance than MFCC features. BirdNet model and the custom CNN models both make use of spectrogram features and seem to be the best models that are comparable.

It can be observed from the table that overall the best-performing model is the custom CNN model using the spectrogram features which shows good performance across all metrics. BirdNet model has a similar performance with a slightly reduced precision and f1-score. CNN model using MFCC fea-

Hyperparameter	Range of Values
Number of CNN-layers	[2 to 5]
Optimizer	['adam', 'rmsprop', 'sgd']
Learning Rate	[0.00001, 0.0001, 0.01, 0.1]
Batch Size	[32, 64, 128]
Epochs	[10, 50, 400]
Dropout Rate	[0.0, 0.25, 0.33, 0.5, 0.75, 0.9]
MaxPool2D Pool Sizes	[(2, 2), (3, 3)]
Conv2D Filter Sizes	[(3, 3), (5, 5)]
Dense Units	[512, 1024, 2048]

Table 1: *Hyperparameters optimised*

Table 2: *Noise Reduction and Data Balancing Experiments*

Model	Technique	Data	NR	Acc(%)
CNN	MFCC	Unbalanced	✓	96.23
CNN	MFCC	Balanced	✓	95.51
CNN	MFCC	Unbalanced	✗	90.11
CNN	MFCC	Balanced	✗	91.01

tures (Fig 4) is comparable to the InceptionV3 and Xception models for all error metrics, with slightly lower recall scores. ResNet50 spectrogram models and the LSTM-based RNN models using MFCC features are both low in recall and f1-score (Fig 5) compared to other models. While the ResNet50 model seems to have a very high precision score. This model could be overfitting and hence unable to extract all the Jackdaw calls in the validation set. The transformer model using the raw waveforms (using wav2vec) also does not match the performance of the custom CNN and BirdNet models but has consistent scores across all metrics including accuracy, precision, recall, and f1-score(Fig 6). More fine-tuning and optimisation might be needed to improve the transformer model performance.

The detection accuracy of the model is manually validated by an ornithologist, an expert researcher familiar with western jackdaw vocalisations. The best-performing fine-tuned BirdNet model could extract 471 Jackdaw calls from the 17.7 hours of raw unlabeled field test data mentioned earlier. The predictions are filtered with a threshold confidence score above 50% to gather around 248 Jackdaw calls to be manually validated to understand the precision of the model on a noisy real-world dataset. It was observed that almost 230+ clips were observed to contain Jackdaw calls resulting in around 93% precision. The confusing sounds in the wrongly extracted audio were mostly noise, a strong gush of wind. In the future, better models will be trained with more false-positive sounds added to the background model to avoid wrong triggers. Also, appropriate confidence threshold values will be experimented with to give the most appropriate segments of Jackdaw vocalisations, considering the balance between missed detections and false positives.

4. Conclusion

In conclusion, our study addresses the challenge of automated detection and classification of western jackdaw calls, essential for population monitoring and behavioral research across Europe. Through a convolutional neural network (CNN) approach robust to environmental noise, we achieve over 98% validation accuracy on a curated dataset of jackdaw calls and noise segments from field recordings. Multiple models are being evaluated using MFCC and spectrogram features. The deep learning models studied include custom CNN models and multiple pre-trained models, BirdNet, ResNet50, InceptionV3, and Xception models. LSTM-based sequence models and pre-trained transformer models are also experimented with. The best-performing model on the labeled test data was the BirdNet model or the custom CNN model using the spectrogram features. Employing data balancing and aggressive noise filtering, our model demonstrates improved generalization under noisy conditions. This advancement marks progress in automated avian census and monitoring, with potential as a benchmark for future studies in avian audio classification.

Table 3: *Comparing multiple deep learning models*

Model	Feature	Acc(%)	P(%)	R(%)	F1
CNN	MFCC	96.23	94.91	87.23	92.10
BirdNet	Spectrogram	98.31	97.94	98.03	97.98
CNN	Spectrogram	98.91	98.90	98.93	98.91
ResNet50	Spectrogram	89.45	1.0	65.51	79.16
InceptionV3	Spectrogram	95.36	98.49	88.43	93.19
Xception	Spectrogram	93.67	93.39	89.01	91.39
LSTM	MFCC	90.00	80.02	77.20	78.60
Transformers	Wav2Vec2	89.31	89.21	89.33	89.31

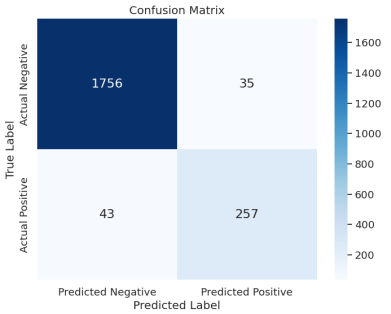

Figure 4: *CNN MFCC Confusion Matrix*

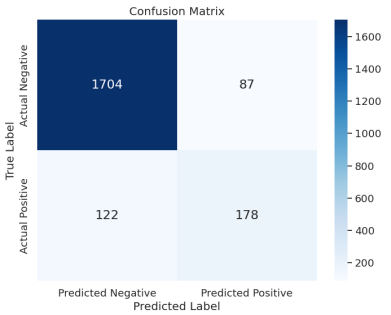

Figure 5: *LSTM MFCC Confusion Matrix*

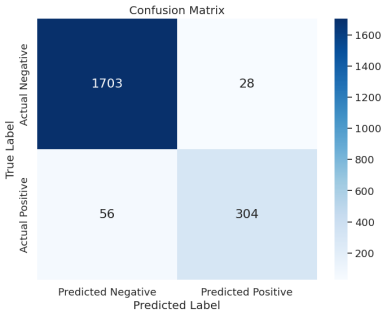

Figure 6: *Wav2Vec2 Transformer Confusion Matrix*

5. References

[1] S. D. H. Permana, G. Saputra, B. Arifitama, W. Caesarendra, R. Rahim *et al.*, "Classification of bird sounds as an early warning method of forest fires using convolutional neural network (cnn) algorithm," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 7, pp. 4345–4357, 2022.

[2] S. Kahl, C. M. Wood, M. Eibl, and H. Klinck, "Birdnet: A deep learning solution for avian diversity monitoring," *Ecological Informatics*, vol. 61, p. 101236, 2021.

[3] J. Stastny, M. Munk, and L. Juranek, "Automatic bird species recognition based on birds vocalization," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2018, no. 1, pp. 1–7, 2018.

[4] A. Incze, H.-B. Jancsó, Z. Szilágyi, A. Farkas, and C. Sulyok, "Bird sound recognition using a convolutional neural network," in *2018 IEEE 16th international symposium on intelligent systems and informatics (SISY)*. IEEE, 2018, pp. 000 295–000 300.

[5] E. Cakir, S. Adavanne, G. Parascandolo, K. Drossos, and T. Virtanen, "Convolutional recurrent neural networks for bird audio detection," in *2017 25th European signal processing conference (EUSIPCO)*. IEEE, 2017, pp. 1744–1748.

[6] J. Xie, K. Hu, M. Zhu, J. Yu, and Q. Zhu, "Investigation of different cnn-based models for improved bird sound classification," *IEEE Access*, vol. 7, pp. 175 353–175 361, 2019.

[7] E. Sprengel, M. Jaggi, Y. Kilcher, and T. Hofmann, "Audio-based bird species identification using deep learning techniques," *LifeCLEF 2016*, pp. 547–559, 2016.

[8] M. Sankupellay and D. Konovalov, "Bird call recognition using deep convolutional neural network, resnet-50," in *Proc. Acoustics*, vol. 7, no. 2018, 2018, pp. 1–8.

[9] I. Nolasco, S. Singh, V. Morfi, V. Lostanlen, A. Strandburg-Peshkin, E. Vidaña-Vila, L. Gill, H. Pamuła, H. Whitehead, I. Kiskin, F. H. Jensen, J. Morford, M. G. Emmerson, E. Versace, E. Grout, H. Liu, B. Ghani, and D. Stowell, "Learning to detect an animal sound from five examples," *Ecological Informatics*, vol. 77, p. 102258, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S157495412300287X

[10] J. Benesty, J. Chen, and Y. A. Huang, "Noise reduction algorithms in a generalized transform domain," *IEEE transactions on audio, speech, and language processing*, vol. 17, no. 6, pp. 1109–1123, 2009.

[11] Z. Zhao, L. Yang, R.-r. Ju, L. Chen, and Z.-y. Xu, "Acoustic bird species classification under low snr and small-scale dataset conditions," *Applied Acoustics*, vol. 214, p. 109670, 2023.

[12] T. Grill and J. Schlüter, "Two convolutional neural networks for bird detection in audio signals," in *2017 25th European Signal Processing Conference (EUSIPCO)*. IEEE, 2017, pp. 1764–1768.

[13] S. BN, "Automatic bird sound detection in long range field recordings using wavelets & mel filter bank features," in *2020 IEEE Second International Conference on Cognitive Machine Intelligence (CogMI)*, 2020, pp. 218–226.

[14] W. Ansar, A. Chatterjee, S. Goswami, and A. Chakrabarti, "An efficientnet-based ensemble for bird-call recognition with enhanced noise reduction," *SN Computer Science*, vol. 5, no. 2, p. 265, 2024.

[15] C. Srujana, B. Sriya, S. Divya, S. Shaik, and V. Kakulapati, "Species identification of birds via acoustic processing signals using recurrent network analysis (rnn)," in *Soft Computing and Signal Processing*, H. Zen, N. M. Dasari, Y. M. Latha, and S. S. Rao, Eds. Singapore: Springer Nature Singapore, 2024, pp. 27–38.

[16] A. Noumida and R. Rajan, "Multi-label bird species classification from audio recordings using attention framework," *Applied Acoustics*, vol. 197, p. 108901, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0003682X22002754

[17] Y. Zhang and J. Li, "Birdsoundsdenoising: Deep visual audio denoising for bird sounds," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2023, pp. 2248–2257.

[18] A. P. Hill, P. Prince, J. L. Snaddon, C. P. Doncaster, and A. Rogers, "Audiomoth: A low-cost acoustic device for monitoring biodiversity and the environment," *HardwareX*, vol. 6, p. e00073, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2468067219300306