Neurocomputing 634 (2025) 129892

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Learning with noisy labels for classifying biological echoes in polarimetric weather radar observations using artificial neural networks

John Atanbori [a,*], Christos A. Frantzidis [a], Mohammed Al-Khafajiy [a], Aliyu Aliyu [a], Behnaz Sohani [b], Kofi Appiah [c], Harriet Moore [d], Catherine Sanders [e], Alastair I. Ward [f]

[a] University of Lincoln, School of Engineering and Physical Sciences, Brayford Way, Brayford Pool, Lincoln, LN6 7TS, United Kingdom
[b] Loughborough University, Wolfson School of Mechanical, Electrical and Manufacturing Engineering, Epinal Way, Loughborough, LE11 3TU, United Kingdom
[c] University of York, Department of Computer Science, Deramore Lane, York, YO10 5GH, United Kingdom
[d] University of Lincoln, Lincoln Institute for Rural and Coastal Health, Brayford Way, Brayford Pool, Lincoln, LN6 7TS, United Kingdom
[e] University of Lincoln, School of Natural Sciences, Brayford Way, Brayford Pool, Lincoln, LN6 7TS, United Kingdom
[f] University of Leeds, School of Biology, Leeds, LS2 9JT, United Kingdom

ARTICLE INFO

Communicated by Y. Bao

Keywords:
Artificial neural networks (ANN)
Ensemble classifiers
Radar bio-scatterer classification
Semi-supervised co-training

ABSTRACT

The identification of biological echoes in radar data has revolutionized research into airborne migratory species. Deep learning applied to polarimetric weather radar observations can reveal signature patterns of mass movement by bio-scatterers such as birds, bats, and insects. However, due to the difficulties in labelling bio-scatterers in these data, threshold approaches have been proposed in the literature. In this research, we used the depolarization ratio (DR) based on differential reflectivity (zDR) and the cross-correlation coefficient (pHV), along with citizen scientist-reported data, to label bio-scatterers for deep learning. This method of labelling biological echoes in radar signatures is prone to noise, which impacts the accuracy of any model that relies on it. We introduce a novel semi-supervised co-training approach that uses a bootstrap ensemble with a confidence threshold. Our ensemble consists of the newly proposed STNet and two modified FNet models, which incorporate co-learning through bootstrap sampling for label correction. This innovative method significantly improves classification accuracy across all three multivariate numerical datasets compared to baseline models that lack co-learning with bootstrap-based label correction.

1. Introduction

Monitoring population trends is a fundamental component of species conservation and management. It is of growing importance as human impacts increase the necessity for conservation management of wild populations [1]. The UN Convention on the Conservation of Migratory Species of Wild Animals (also known as the Bonn Convention) recognizes the importance of migratory species and hence their conservation (www.cms.int). However, some migratory species are also known to facilitate the inter-continental spread of infectious diseases that threaten human interests, such as avian influenza [2]. In addition to their epidemiological significance, migratory species play crucial ecological roles, such as seed dispersal, pollination, and nutrient cycling. Therefore, accurate monitoring methods are essential not only for conservation purposes but also for understanding the broader ecological impacts of these species.

Monitoring species over a large spatial scale is daunting, but polarimetric weather radar can simplify the process as they are geographically distributed across multiple regions worldwide. Weather radar systems, initially developed for meteorological applications, have become invaluable tools in ecological research. Their ability to provide continuous, large-scale observations without direct disturbance to wildlife offers a unique advantage over traditional field survey methods. This advancement supports diverse applications, from migration tracking to the study of population dynamics.

Polarimetric weather radar measurements have been used to separate meteorological and non-meteorological scatterers [3,4]. They measure six single- and dual-polarization variables. The single-polarization variables consist of radar reflectivity factor (Z), velocity (V), and spectrum width (SW), traditionally used in weather applications for removing non-weather echoes. The introduction of dual-polarization variables: differential reflectivity (zDR), cross-correlation coefficient (rHV), and Differential Phase (PH) have led to improved algorithms for meteorological and non-meteorological applications [3–6]. In particular, the discrimination of Biological scatterers in non-meteorological echoes [5–8], used in the detection of bird roosts [8], quantifying species

* Corresponding author.
E-mail address: jatanbori@lincoln.ac.uk (J. Atanbori).

https://doi.org/10.1016/j.neucom.2025.129892
Received 10 April 2024; Received in revised form 17 February 2025; Accepted 1 March 2025
Available online 10 March 2025

emerging from roosts [9], and classification of biological targets [5]. Many applications using weather radar data classify scatterers into three categories, namely precipitation, biology, and clutter. These applications [5–7,9] often rely on the distinctive properties of combined single- and dual-polarization variables. As these radar systems are operational across most developed nations, their data hold the potential to be a cost-effective global solution for tracking species at risk and mitigating ecological threats.

Recently, machine and deep-learning approaches have been used with applications that use polarimetric weather radar measurements [5,8,10–12]. However, deep learning usually requires a large dataset of radar scans with labels, but these data are usually difficult to annotate. Lin et al. [10] used transfer learning from image classification models trained on the ImageNet dataset [13] to overcome the problem, and classified radar echoes into biological and non-biological classes. Manually, labelling echoes in polarimetric radar data could be very time-consuming and error-prone. Therefore, research has used thresholding [4,10] to speed up radar scatterers labelling. The two most common thresholding approaches include the cross-correlation coefficient used in [7] to identify biological scatterers and the depolarization ratio used in [4] to discriminate non-biology signatures from biology. Approaches that use citizen science (CS) for species-specific labelling have also been proposed [5,14], but these are based on assumption of clear-air (no precipitation) and minimal scatterer cross-contamination. However, mixed-scatterer types in radar signature is common and some degree of cross-contamination from non-focal scatterers are inevitable. Therefore, these approaches introduce noisy labels [5] that affect model training and introduce errors into the prediction of bio-scatterers echoes. In this study, we focused on noisy labels in general rather than those specifically introduced by cross-contamination, as the issue is not only limited to this.

The noisy label problem is a rapidly emerging research theme in the deep learning community. One approach to addressing the problem involves selecting correctly labelled examples from noisy training datasets by eliminating labels likely to be mislabelled in order to ensure robust learning [15]. This approach then combines the clean and noisy labels (as unlabelled data) and uses semi-supervised learning to improve the model's predictive accuracy. Another approach relies on a loss correction strategy, which is used to reduce the impact of noisy labels during the network training stage by directly modifying (or adjusting) the losses through various methods [16]. In contrast to conventional approaches, our method employs a semi-supervised co-training strategy, utilizing a bootstrap ensemble with a confidence threshold for annotating noisy labels. Additionally, we incorporate a bootstrap sparse categorical cross-entropy loss to enhance the classification. The advantage of our approach lies in the collaborative training of ensemble models with distinct data subsets, facilitating information sharing between them. This synergistic approach enhances overall performance by harnessing the predictive power inherent in ensemble learning, co-training and the bootstrap sparse categorical cross-entropy loss. The contribution of this paper is as follows:

- Label biological echoes in polarimetric weather radar observation using depolarization ratio [4] and citizen science data.
- Attempt noisy biological-scatterers label correction in polarimetric weather radar using a deep learning approach.
- Introducing our novel Short-Time Fourier Transform network (STNet), designed to compute the FFT of short overlapping segments of the input while incorporating temporal information for our ensembles.
- Furthermore, we have integrated ensembles with a combination of bootstrap sparse categorical cross-entropy and a co-training approach to address noisy labelling. This innovative combination has proven effective in classifying biological-scatterer signatures in polarimetric weather radar data, even in the presence of label noise.

The remainder of this paper is organized as follows. We review existing deep learning literature that learns from noisy labels in Section 2. Section 3 describes our methods. Then we move on to Section 4 to describe the datasets, experiments, and benchmarking. Finally, in Section 5, we present and discuss the results, and in Section 6, we conclude.

2. Related work

2.1. Discriminating biological echoes

To date, very little work exist on discriminating biological scatterers in polarimetric weather radar using Deep learning approaches [10–12]. Usually, massive amounts of labelled data are required to train these models, which is time-consuming and expensive to obtain from weather radar observations. Convolutional neural networks were employed by [10–12] to discriminate between biological and non-biological scatterers; however, these methods necessitate extensive datasets for training. To avoid the cost and time limitations, [10] used a transferred learning approach which is based on the IMAGENET dataset [13] with only a few manually labelled radar data. The weak labels were created using a cross-correlation threshold of 0.95, a common practice among radar biologists to identify biological echoes in radar data. However, [10] recommended that the use of Depolarization Ration(DR) [4] could be a better alternative to their approach for labelling the radar data. Labelling biological echoes this way introduces cross-contamination with non-focal scatterers. Gauthreaux et al. [5] opted not to label a significant amount of radar data as they sought to discriminate among six different types of biological scatterers using natural history and the Random Forest (RF) approach. Nevertheless, the method remains susceptible to non-focal scatterers, possibly due to the lack of independent validation for scatterer types.

Research studies on roost detection using deep learning techniques, as highlighted in the literature [8,17], have demonstrated a reliance on a larger volume of labe1led data compared to pixel-level approaches [10–12,12]. This preference is attributed to the convenience of delineating bounding boxes around distinctive ring-roost patterns. The characteristic nature of these patterns is particularly evident in the emergence of tree swallow roosts. Nonetheless, it is crucial to acknowledge that such patterns may not be readily discernible for other species, emphasizing the need for a non-species-specific approach when employing deep learning methods for diverse roost detection scenarios. Our approach focuses on the mass migration of bird species, and it is not dependent on ring-roost pattern and aims to find a solution that is not species-specific. While the conventional approaches emphasis on recognizing characteristic roost patterns in this case, our method shifts the paradigm to address the complexities of large-scale migration events, which could broaden the understanding of roost detection dynamics for varying species. However, this study only focused on large-scale migration footprints.

2.2. Noisy labels techniques

Polarimetric weather radar data typically has no or very little labelled data, but supervised deep learning requires massive amounts of labelled data. Although there are techniques to quickly annotate these data, they introduce noisy labels since mixed scatterers are unavoidable and difficult to separate manually. Supervised Learning models for classification using noisy data will inevitably degrade during training [18]. In this section, we have followed a similar categorization approach as in [15,16] to review the literature in this area. However, because this is not a survey paper, we strongly advise readers who want more information to read the articles by [15,16]. Furthermore, the effectiveness of similarity relationship hashing for unsupervised cross-modal retrieval has been explored in recent research [19]. These findings could potentially contribute to improving the noisy label classification problem by leveraging cross-modal relationships to refine data labelling strategies.

2.2.1. Loss correction methods

Some noisy label technique aims to correct the training loss [16, 20–22] by building a regularization into the loss function to penalize low confident predictions. Hendrycks et al. [21] and Patrini et al. [22] employed a noise transition matrix, which signifies the transition relationship from clean labels to noisy ones, to construct statistically consistent classifiers in label-noise learning. This approach may lead to a poorly estimated transition matrix due to the randomness of label noise, a problem addressed in [20] using two easy-to-estimate transition matrices, known as a dual-T estimator. The commonly used cross-entropy loss is not robust to noisy labels. Therefore, loss functions such as information-theoretic loss [23] and normalization loss [24], which mathematically lower the impact of noisy labels during back-propagation have recently been proposed. Robust approaches could also include designing specific loss layers [25], which can improve network learning by correcting the noisy labels.

Another popular loss correction approach corrects labels using a label propagation before training the models on the pseudo (corrected) labels since the noisy labelled data have the same feature distribution as the clean data [26]. Though [27] is based on pseudo labels, they treat corrected labels as an independent parameter learned during training. However, incrementally correcting the training data in each epoch and updating the model with pseudo labels could lead to the network memorizing the noisy labels, which [28] address by adding a regularization on the loss function to lower the possibility of incorrect predictions. Our approach is partially inspired by Liu et al. [28], where we conduct the label propagation process to acquire pseudo labels within a semi-supervised co-training network, leading to improved model predictions. Additional details about this model are provided in Section 3.

2.2.2. Sample selection methods

Both loss correction and sample selection methods modify the loss by reducing the harmful effects of noisy labels during training. However, the latter divides the data into clean and noisy sets using a Gaussian mixture model, or small loss criterion in the case of co-teaching models [29–31]. The sample selection method uses clean data for training. Early sample selection methods proposed in [29] train two deep neural networks simultaneously and let them teach each other given every mini-batch. The approach feed-forward all the training data and selects some data of possibly clean labels. The two networks then jointly decide which mini-batch is used for subsequent training. Mandal et al. [30] and Wei et al. [31] proposed improvements to the co-teaching approach. Mandal et al. [30] incorporated modifications such as self-supervision and relabelling into the co-teaching framework. Conversely, Wei et al. [31], following a strategy similar to that of Han et al. [29], chose to select small-loss examples for simultaneously updating the parameters of both networks. In alignment with the methodologies explored in this section, our approach introduces an ensemble comprising a Short-Time Fourier Transform network (STNet) and a Fast Fourier Transform network (FNet). This ensemble employs a hybrid strategy, combining bootstrap sparse categorical cross-entropy and a co-training approach, to effectively tackle the challenges posed by noisy labelling.

3. Methods

Our goal is to employ a method that minimizes the cross-contamination of labelled scatterers in polarimetric radar data and utilizes a deep learning approach to reduce the propagation of noisy labels throughout the network. The methods employed to achieve this are detailed in this section.

3.1. Depolarization ratio

Since we do not have a large labelled dataset of polarimetric observations, we used one of the popular approaches, Depolarization Ratio to first separate radar signatures into metrological and non-metrological. We then used citizen data and existing sightings of species to separate species specific data from the non-metrological data to produce a comprehensive dataset for training our models. While it is a common practice among radar biologists to use a threshold of $pHV \leq 0.95$, as employed in prior work for weakly labelling data in deep learning [10], we opted for the Depolarization Ratio as recommended in [10]. This choice is anticipated to yield better results, and the computation is performed using Eq. (1).

$$DR = 10\log_{10}\left(\frac{zDR + 1 - 2(pHV)\sqrt{zDR}}{zDR + 1 + 2(pHV)\sqrt{zDR}}\right) \tag{1}$$

where zDR and pHV are the differential reflectivity and cross-correlation products respectively in linear scale. The unit of measurement for Differential Reflectivity (zDR) is decibels (dB). It measures the ratio of horizontally polarized reflectivity to vertically polarized reflectivity, with a typical value range of -7dB to $+7,$dB. Meanwhile, the Cross-Correlation Coefficient (pHV) is dimensionless (unitless). It represents a normalized measure of the similarity or correlation between horizontally and vertically polarized radar returns, with values ranging from 0(not correlated) to 1(perfect correlation). Our computation of DR has been converted into decibels (dB) but usually its values range from 0 to 1. The DR of meteorological targets is small, but hail and melting graupel could have values of DR as high as non-meteorological targets, but have reflectivity (Z) never observed in biological echoes [4]. We, therefore, used $DR \leq -12$ and $Z \leq 40$ similar to those proposed in [4,32] to label biological echoes.

3.2. Semi-supervised co-training with bootstrap ensemble

Fig. 1 demonstrates our semi-supervised co-training method, incorporating a bootstrap ensemble and a confidence threshold. We annotate our unlabe1led training data ($D_{\text{unlabelled}}$) by leveraging the best models obtained through iterative ensemble training. The application of a confidence threshold (θ) ensures that only predictions (x) surpassing this value contribute to the annotations. We then update the complete training dataset with these confident predictions, forming a refreshed dataset (see D_{train} in Eq. (2)) for use in the next co-training iteration.

$$D_{\text{train}} \leftarrow D_{\text{train}} \cup \{(x, \text{Label}(M_i, x)) \mid x \in D_{\text{unlabe1led}}, \\ \text{Confidence}(M_i, x) > \theta\} \tag{2}$$

This iterative cycle continues as the ensemble training process recommences, utilizing the updated training data. The iterative refinement contributes to the continual improvement of the models' accuracy. The iterative process persists until the best ensemble models' predictive accuracy plateaus, and they cannot predict any more unlabe1led data with the required confidence threshold, signifying a convergence of the ensemble models (see Eq. (3)).

$$\text{Confidence}(M_{\text{best}}, x) < \theta, \forall x \in D_{\text{unlabe1led}} \tag{3}$$

Following convergence, we undertake a final training using the updated dataset (see Eq. (4)), which focuses on STNet, our novel architecture (see 1). However, the final model could also be the FNet or BasicFNet models, but we selected STNet as it provides better accuracy overall.

$$\text{Final Model} = \text{Train}(D_{\text{train}}, \text{STNet}) \tag{4}$$

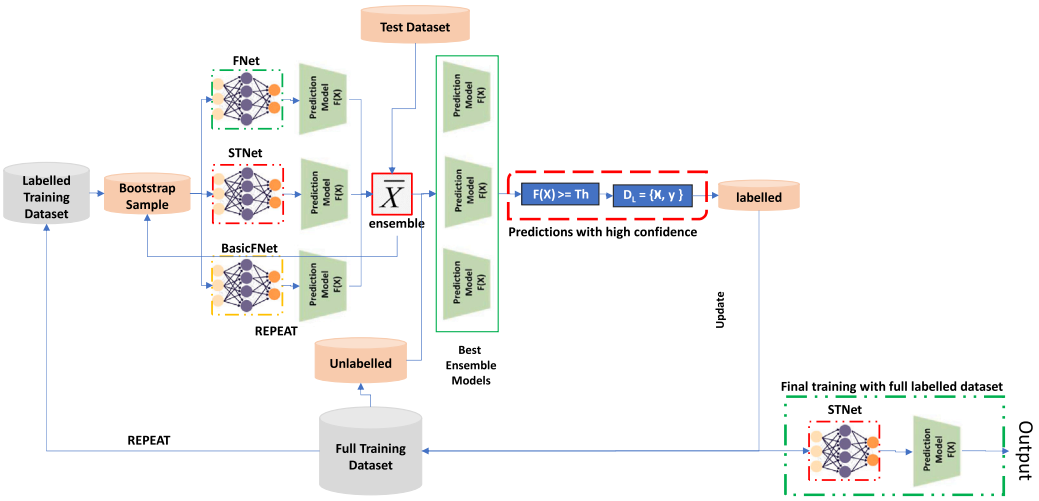

Fig. 1. The semi-supervised co-training approach introduces a novel bootstrap ensemble that aggregates results from multiple models – FNet, STNet, and BasicFNet – each contributing unique modelling strategies for improved prediction performance. FNet is designed to capture feature-based dependencies, while STNet leverages spatiotemporal information, and BasicFNet serves as a baseline with a simpler architecture. A key innovation is the use of a confidence threshold, which selectively incorporates only reliable predictions for labelling, preventing the influence of noisy or uncertain data. The training iterates until further labelling no longer improves performance, ensuring efficient use of labelling only when it benefits the model's refinement.

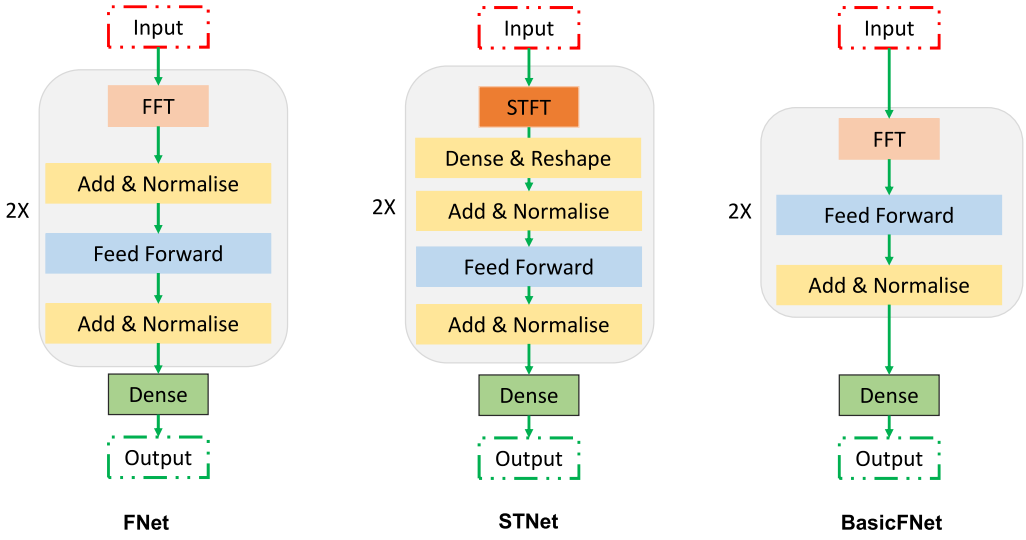

Fig. 2. Architecture of the Bootstrap Ensembles: The FNet architecture (left) is based on the design proposed in [33], with a modification where the FFT layer is 1D instead of 2D. The sequence of layers from the FFT layer to the second Add & Normalized block is repeated multiple times before passing through the Dense and output layers. The STNet architecture (middle) follows a similar structure to FNet but replaces the FFT layer with an STFT layer, followed by Dense and Reshape layers. Finally, the BasicFNet architecture is a simplified version of FNet, omitting the initial Add & Normalized layer.

3.3. The architectures for bootstrap ensemble learning

We employed three ensembles, namely FNet, STNet, and BasicFNet, in the context of semi-supervised Co-Training. FNet serves as an attention-free alternative to conventional Transformer architectures, as discussed in [33]. Our objective is to utilize models with reduced computational complexity to shorten training time while still delivering superior performance. This is achieved by replacing the self-attention mechanism with a more efficient component. The FNet architecture we employ, shown in Fig. 2a, is based on the implementation in [33] but introduces a key modification in the Fourier sublayer. Unlike [33], which uses a 2D FFT, our approach implements a 1D FFT, specifically designed to better handle the characteristics of the multivariate datasets in this study. Additionally, we used the real components of the FFT, following the recommendations demonstrated in [33].

The FNet architecture comprises two major sublayers: the Fourier Mixing Sublayer and the Feed-Forward Sublayer. The Fourier Mixing Sublayer is responsible for capturing dependencies in the input sequence by transforming it into its frequency domain. Given our specific use case with short input sequences, we padded the input to a total length of 256. Meanwhile, the Feed-Forward Sublayer is a standard feed-forward neural network sublayer designed to process the outputs of the FFT.

Fourier Mixing Sublayer: This sublayer uses a mathematical tool called the *Fourier Transform*, which is a method for analysing signals (or sequences of data) by breaking them down into their frequency components. Think of it like taking a piece of music and identifying the individual notes and harmonies that make up the song. Similarly, in the Fourier Mixing Sublayer, the input sequence (a series of numbers representing text, speech, or other data) is transformed into its *frequency domain*. This allows the model to uncover patterns or relationships in the data that may not be obvious in its original form, such as recurring structures or long-range dependencies. To make this work efficiently, the model ensures that all input sequences are of the same length,

which in this case is *256*. If the input is shorter than this, it is "padded" by adding extra values to reach the desired length. This padding does not add new information but ensures consistency for the transformation and subsequent steps.

Feed-Forward Sublayer: Once the data has been processed in the Fourier Mixing Sublayer, the resulting frequency information is passed to the *Feed-Forward Sublayer*. This part is a type of standard *neural network layer*. It takes the transformed data and applies a series of computations designed to extract meaningful features and patterns. Essentially, this layer helps the model make sense of the frequency-domain data and prepares it for the final stages of processing, such as making predictions or classifications.

We introduce STNet, a novel adaptation of the FNet architecture, which incorporates the Short-Time Fourier Transform (STFT) for enhanced time–frequency localization. This key innovation allows STNet to effectively capture variations in the input signal across both time and frequency domains. Unlike the FFT sublayer in FNet, STNet features a unique STFT sublayer, followed by dense and reshape layers (see Fig. 2b). To accommodate this architecture, we implemented an STFT operation with a frame length of 256 and a step of 2, producing an output shape that required reshaping for compatibility with the network's structure. This was achieved using a dense layer with 256 units, ensuring seamless integration within STNet.

We introduce BasicFNet, a novel simplified version of the FNet architecture (see Fig. 2c). In this variant, we omit the Add and Normalized layers that typically follow the FFT sublayer in the original FNet design. This deliberate modification creates diversity in the ensemble's predicted outputs, which is crucial for enhancing the co-training process. By intentionally introducing variation among the ensemble members, BasicFNet improves the overall robustness and effectiveness of the co-training mechanism.

3.3.1. Architecture implementation details

The three architectures - **FNet**, **BasicFNet**, and **STNet** - are closely related. Each model applies a transformation (FFT for **FNet** and **BasicFNet**, STFT for **STNet**) to the input, followed by layer normalization with $epsilon = 1e{-}6$ to stabilize the learning process. A feed-forward network, consisting of a dense layer with 256 neurons and Gaussian Error Linear Unit (GELU) activation, refines the features in all architectures. Skip connections are used to add the output of the feed-forward layer back to the transformed input, followed by another layer normalization. The final output layer in each architecture uses softmax activation to produce class probabilities. All three models are compiled with the Adam optimizer and a custom sparse categorical loss function based on bootstrapping (see Section 3.4).

- **FNet:** The architecture repeats the FNetLayer, where each layer applies FFT to the input (cast to float32), followed by the components mentioned above.
- **BasicFNet:**The BasicFNetLayer applies FFT to the input, followed by a similar sequence of transformations and refinements as the FNet, except that there is no skip connection immediately after the FFT operation.
- **STNet:** The STNetLayer applies a Short-Time Fourier Transform (STFT) to the input with $frame\_length = 256$, $frame\_step = 2$, and $fft\_length = 256$. The rest of the architecture, including feed-forward refinement, skip connections, and final softmax output, follows the same structure as FNet and BasicFNet.

These codes are available open-source at the GitHub Repository https://github.com/Amotica/RadMLProofvFinal

3.4. The bootstrap sparse categorical cross entropy loss

Bootstrapping loss is a technique that leverages the inherent uncertainty in the model's predictions to mitigate the impact of noisy labels [34]. In simpler terms, it adjusts how much importance is given to each training sample based on how confident the model is in its predictions, reducing the negative effect of incorrect labels. We implemented a variant of sparse categorical cross-entropy known as Bootstrap Sparse Categorical Cross-Entropy, which assigns different weights to the training samples based on their prediction confidence. This means that samples the model is less confident about contribute less to the loss calculation, helping the model focus on cleaner, more reliable data.

This loss function introduces a novel approach by incorporating the Bootstrap Sparse Categorical Cross-Entropy Loss, which multiplies the Sparse Categorical Cross-Entropy loss by the weight of the true class labels ($W_i \cdot y_i$). The key innovation here lies in the application of these weights, which effectively modulate the learning process by placing greater emphasis on more reliable labels while down-weighting those that are noisy or uncertain. These weights essentially act as a dynamic filter, guiding the model to focus on trustworthy data and reducing the influence of questionable labels. Each of the three models in the ensemble employs this loss function, which not only mitigates the impact of noisy labels but also strengthens the model's overall robustness. The Bootstrap Sparse Categorical Cross-Entropy Loss is formally defined in Eq. (5).

$$-\sum_i \mathrm{W}_i \cdot y_i \cdot \log\left(\frac{e^{z\bar{y}_i}}{\sum_j e^{\bar{y}_j}}\right) \tag{5}$$

where:

i represents index of the classes.

W_i denotes the weight assigned to each sample for class i.

y_i is the true label for class i.

z is score for the true class.

$\bar{y}_i$ is the predicted probability for class i.

The weights (W_i) in this loss function normalizes each sample's contribution based on batch size and bootstrap samples, ensuring a balanced impact proportional to positive labels. This normalization step ensures that no individual class or sample disproportionately influences the learning process. The Bootstrap Sparse Categorical Cross-Entropy Loss uses these weights in computing the loss for each batch sample, considering log-likelihood with respect to true labels, and is calculated as in Eq. (6).

$$\mathrm{W} = \frac{\sum_i y_i}{b \times n} \tag{6}$$

where:

$\sum_i y_i$ is the sum of true labels across all classes.

b is the batch size.

n is the number of bootstrap samples.

In summary, the Bootstrap Sparse Categorical Cross-Entropy Loss dynamically adjusts how much weight each sample has on training, based on prediction confidence and label quality. This makes it a powerful tool for training robust models, especially in datasets where label noise is a concern.

3.5. The confidence threshold

A key novel contribution of this work is the introduction of a confidence threshold within the context of co-training. This threshold establishes a predefined level of certainty or confidence, acting as a filter for predictions generated by the models during the co-training

process. It plays a crucial role in determining which predictions are considered reliable and subsequently used to update the training dataset. Only predictions that surpass the confidence threshold are incorporated into the training set, actively enhancing the iterative refinement of the models throughout the co-training procedure.

In our case, the confidence threshold, denoted as (C) and illustrated in Eq. (7), introduces a mechanism for customizing the level of confidence through the integration of a user-defined parameter called the confidence delta (δ). The confidence δ is a user-defined parameter that adjusts the confidence threshold, allowing flexibility in controlling how confident a prediction must be to be included in the training process. The confidence delta is a flexible parameter that users can adjust to fine-tune the confidence threshold according to specific requirements. This enables tailoring the confidence level to meet diverse scenarios or align with different characteristics datasets, enhancing the practicality and effectiveness of the co-training approach.

$$C = \frac{1}{N} + \delta \tag{7}$$

where:

C is the confidence threshold,

N is the number of classes,

δ is the confidence delta

$\frac{1}{N}$ is the maximum winning probability

The value of δ directly impacts the model's performance. A higher δ results in a stricter threshold, making the model more conservative by only accepting highly confident predictions. This reduces the risk of errors but may slow down learning. When the confidence delta (δ) is increased, the required level of certainty for predictions to be included in the training set is raised. This means only highly confident predictions (those with a probability exceeding the threshold) are added to the training set. While this helps reduce the likelihood of incorrect or uncertain predictions being incorporated into the training data, it also limits the number of predictions available for training. As a result, the model has fewer opportunities to refine itself, which slows down the learning process. Because of this, δ should be obtained via grid search or manual continuous experimentation to find a good value that works for your specific dataset. In all conducted experiments, we consistently used a confidence delta value of 0.4, determined through experimentation with our datasets. However, it is crucial to highlight that the selection of this value must comply with the constraint that it falls within the range of zero to one minus the maximum winning probability ($0 \leq \delta \leq 1 - \frac{1}{N}$).

4. Datasets and experiments

4.1. Datasets

Our primary emphasis during experimentation centred around the utilization of the Polarimetric weather Radar Dataset. However, to ensure the robustness and versatility of our algorithm, we validated its performance using two supplementary datasets: the DryBeanDataset [35] and the SkinSegmentationRGB [36] datasets. The rationale behind incorporating these additional datasets lies in their shared characteristics as multivariate numerical datasets, closely aligning with the nature of the Polarimetric Weather Radar Dataset. This deliberate choice enables an evaluation of our algorithm across diverse datasets with similar inherent complexities.

4.1.1. Polarimetric radar dataset

The dataset used in the study incorporates NEXRAD Level 2 data obtained from the US Radar network. This data is sourced from the Weather Surveillance Radar-1988 Doppler network operated by the US National Weather Service, comprising approximately 160 radars. These radars conduct volume scans at intervals of around 5 min, performing 360-degree sweeps to capture their data. The dataset has six distinct data products: three legacy products (Z, V, and SW) and three dual polarization products (zDR, DP, and pHV). Dual polarization enhances the radar's capability to discriminate between different types of objects based on shape and uniformity within a pulse volume. In Fig. 3, we present histograms depicting the distribution of six radar variables across the three types of scatterers. Each histogram provides insight into the distribution of each product within the specified classes.

Table 1
Shows the Radar Dataset splits for all six classes used in our experiments by radars, sweeps, and Resolution volumes. We have selected resolution volumes such that this is balanced throughout the classes.

Classes	Radars		Sweeps		Selected RV	
	Train	Test	Train	Test	Train	Test
Bats	5	5	36	19	100,000	25,000
Birds	14	10	37	20	100,000	25,000
Insects	14	9	28	15	100,000	25,000
Other Biology (OBio)	10	8	24	12	100,000	25,000
Weather (WHTR)	9	7	17	9	100,000	25,000
Background (BG)	42	31	118	63	100,000	25,000

In each scan, we initially distinguished between biological and weather echoes using the DR ratio, as outlined in Kilambi et al. [4]. Subsequently, we leveraged information gathered from citizen science, as documented in various literature sources [5,37–41], to further categorize biological entities into three primary classes: birds, insects, and bats. Any biological entities not falling into these specified classes were classified as "other biology", a category encompassing debris, clutter, and similar echoes that do not conform to the criteria for weather based on the DR ratio.

Table 1 displays the scatterer types in the Radar dataset for training and testing. The original dataset had millions of Resolution Volumes (RV) for some classes. However, we implemented a random selection process for resolution volumes to construct the radar dataset. This process aimed to guarantee a fair representation of radars and sweeps, especially resolution volumes. The goal was to ensure a balanced dataset, preventing biased models that might learn specific patterns of the majority class, thus hindering generalization.

4.1.2. Other datasets

To ensure the robustness and versatility of our algorithm, we validated its performance using two supplementary datasets. We selected these datasets due to their multivariate and numerical characteristics, aligning with the nature of the weather radar data under examination. The inclusion of the Dry Bean dataset [35] and the Skin Segmentation dataset [36] ensures a fair validation of the algorithms.

The Dry Bean Dataset [35] comprises 13,611 grain images characterized by 16 extracted features, including 12 dimensions and four distinct shape forms. Meanwhile, the Skin Segmentation dataset [36] consists of 245,057 samples obtained by randomly sampling RGB values from facial images. These images represent diverse demographics and encompass both skin and non-skin samples, with 50,859 samples corresponding to skin and 194,198 representing non-skin entities.

4.2. The experiments

In our evaluation, we employed three datasets, as elaborated in Section 4.1, to conduct a series of experiments. The experiments encompassed the following three sets:

1. **Baseline Experiment:** We initiated the evaluation with a baseline experiment utilizing the `STNet` architecture, devoid of co-training. This experiment specifically utilized the 20% of correctly labe1led and verified datasets.

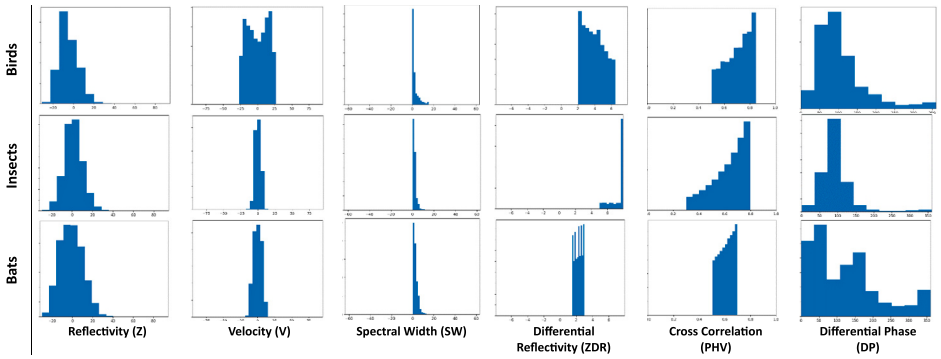

Fig. 3. Illustrates sample histograms for the six radar variables related to the three biological scatterers, showcasing significant overlap in these variables. A subtle, yet discernible difference in characteristics is apparent for the new products (zDR, pHV, and DP). Reflectivity values typically range from −32.0 to 94.5, radar velocity values range from −95.0 to 95.0, spectral width varies between −63.5 and 63.0, differential reflectivity spans from −7.875 to 7.9375, and cross-correlation values are between 0.0 and 1.0. Differential phase values range from 0.0 to 360.0. The provided plots are constrained to these specified ranges.

2. **Semi-supervised Co-training Experiment:** The second set of experiments involved a semi-supervised co-training approach, employing a bootstrap ensemble with a confidence threshold. This method utilized the 20% correctly labe1led and verified datasets in conjunction with the unlabe1led data.
3. **Final Experiment:** The third and final experiment utilized our ultimate models constructed with `STNet`, `FNet`, and `BasicFNet`. The training was performed on the upgraded data resulting from the semi-supervised co-training method based on a bootstrap ensemble with a confidence threshold.

Following this, we performed a comparative analysis between the results obtained from the baseline experiment and those derived from the final model. This comparative assessment aimed to evaluate the effectiveness of the semi-supervised co-training method, which relies on a bootstrap ensemble with a confidence threshold.

4.2.1. Set-up

The semi-supervised co-training method, based on a bootstrap ensemble with a confidence threshold, encompasses several parameters. Some of these parameter values were determined through experimentation with the datasets and may vary depending on the dataset under consideration. We set the confidence delta (δ) value to 0.4 and configured the co-training and bootstrap iterations to 5. The co-training loop is terminated when there is no further improvement in the labelling of the datasets.

Given that a significant portion of the dataset will be unlabelled, we introduced a parameter named “ratio noisy”. We experimented with values of 0.8, 0.7, 0.6, and 0.5 for the proportion of noisy (unlabe1led) data. The fraction of data used as a subset for the ensemble bootstrapping was set at 0.5, determined through experimentation with three datasets. However, this fraction may vary based on the size of the data under consideration, with larger datasets having a smaller fraction.

For all experiments, we standardized the number of epochs to 250 and the batch size to 1024. The objective function for training the network involved a bootstrap sparse categorical cross-entropy loss, and an Adam Optimizer was employed with an initial learning rate of 0.001. The learning rate was then reduced by a factor of ten whenever training plateaued for more than ten epochs.

The CNN models underwent training on a Windows 10 computer equipped with 64 GB of RAM and a 3.6 GHz processor, featuring a GeForce GTX TITAN X GPU with 12 GB of memory. Implementation of all models was carried out using Python 3.6 and Keras 2.3.1 with a TensorFlow backend.

5. Results and discussions

We present the results and discussion in this section. An overarching summary of the outcomes is incorporated within the bar charts illustrated in Fig. 4, with specific attention directed towards the Radar Dataset, recognizing its crucial role as the central nucleus of this research.

Based on the Polarimetric Radar Dataset: Table 2 shows the results of experiments performed using the Polarimetric Radar Dataset. Based on the Baseline experiments, FNet appeared to be the model with the highest accuracy, 85% whereas BasicFNet had the lowest accuracy. This is due to the removal of Add and Normalize layers immediately after the FFT layer, which, although reducing network size, appears to make the model less effective at handling the smaller subset of the dataset obtained via bootstrapping. The collaborative learning approach enhances the precision of all models, highlighting the significant advantages of co-learning in improving model performance on the Radar Dataset. Utilizing ensembles of networks and implementing bootstrap categorical cross-entropy loss enhances the models’ resilience to noisy labels, a crucial factor in accurately labelling the remaining datasets for the final model. It reaffirms assertions in the literature [15,16,18,20] that employing fewer labe1led data or highly noisy labels in supervised learning can lead to overfitting. Our proposed network, STNet, which achieved a classification accuracy of 90%, outperformed other models on this dataset, though it exhibited a lower accuracy (81%) compared with the FNet networks on the baseline experiment. The STNet model was superior because it incorporates temporal and frequency information throughout the learning phase. The inclusion of temporal and frequency layers enhances the model’s capability to handle time-dependent and frequency-specific patterns in data, as it gives it the ability to extract relevant features from data that exhibit temporal dependencies or frequency-specific characteristics.

Analyses using Confusion Matrix: We present the confusion matrix corresponding to the optimal classifier (STNet) in Table 3 to further analysis the methods fit for classifying echoes in the polarimetric radar dataset. In the context of the examined biological echoes, the STNet classifier demonstrated notable accuracy, correctly classifying 93% of insect echoes. However, a significant portion of misclassifications occurred, with the majority erroneously categorized as bat echoes. The classifier achieved an 85% accuracy in identifying bat echoes and a 74% accuracy in discerning bird echoes. Notably, 12% of bird echoes were misclassified as bats, while only 7% of bat echoes were misclassified as birds. The observed similarities between bird and bat echoes, as documented in the literature [32,37], align with our findings. This inherent resemblance contributes to the misclassification patterns encountered in our study.

Weather echoes exhibited minimal misclassification than the biological, with only 2% being mislabelled as birds. We attribute this

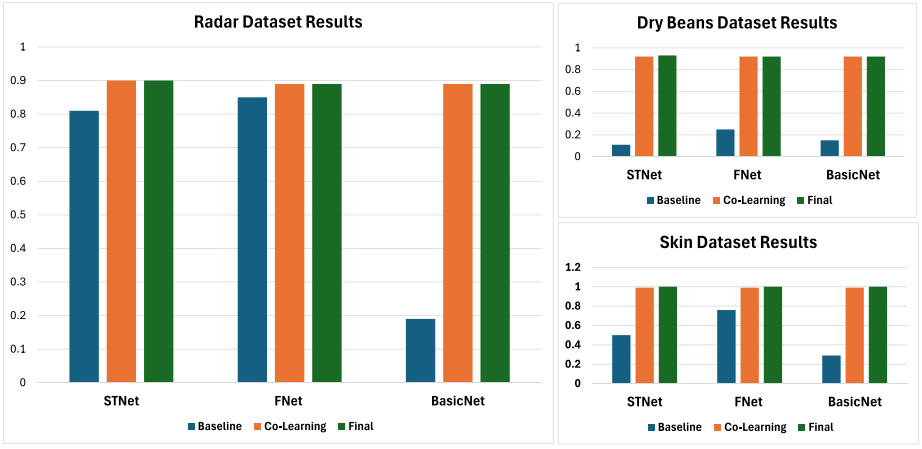

Fig. 4. Summary of the results based on experiments performed using the Radar dataset (left), Dry Beans dataset (top right) and Skin Dataset (bottom right).

Table 2
The results of experiments conducted on the polarimetric radar dataset using the STNet, FNet, and BasicFNet models are presented. The best-performing model is bolded and highlighted in yellow, while the top-performing baseline is italicized and also highlighted in yellow.

	Model	Accuracy	Precision	Recall	F1
STNet	Baseline	0.81	0.82	0.81	0.81
	Co-learning	0.90	0.90	0.90	0.89
	Final Model	**0.90**	**0.90**	**0.90**	**0.90**
FNet	*Baseline*	*0.85*	*0.86*	*0.86*	*0.86*
	Co-learning	0.89	0.89	0.89	0.89
	Final Model	0.89	0.90	0.89	0.89
BasicFNet	Baseline	0.19	0.22	0.19	0.09
	Co-learning	0.89	0.88	0.88	0.88
	Final Model	0.89	0.89	0.89	0.89

Table 3
The STNet Confusion Matrix based on the final model. BG = Background, WHTR = Weather, OBio = Other Biology.

	BG	WHTR	OBio	Bats	Birds	Insects
BG	**0.99**	0.01	0.00	0.00	0.00	0.00
WHTR	0.00	**0.97**	0.01	0.00	0.02	0.00
OBio	0.00	0.01	**0.91**	0.02	0.01	0.05
Bats	0.00	0.00	0.01	**0.85**	0.07	0.07
Birds	0.00	0.01	0.05	0.12	**0.74**	0.08
Insects	0.00	0.00	0.01	0.04	0.02	**0.93**

discrepancy to potential noise in the data labels as a results of automatic labelling using the Depolarization Ratio. The dual polar radar's supplementary products (cross-correlation, differential reflectivity, and differential phase), extensively discussed in existing literature [4,5,10,14], aid in effectively segregating weather echoes. Notably, the non-bird echoes were predominantly classified correctly, highlighting the effectiveness of the proposed approach in handling noisy labels within radar data for such machine learning tasks.

Our suggested approach demonstrated advancements in addressing noisy labels; however, the results indicate opportunities for improving the employed network architecture in the future for biological echoes classification. Suggestions for enhancement include the incorporation of additional labelled data to better differentiate biological echoes or the implementation of post-processing algorithms, such as a despeckling algorithm, which aims to eliminate false alarms in the nearest neighbourhood by assessing whether the label of centre pixel differs from the majority of itself and its eight immediate neighbours [4] an approach that will effectively minimize some misclassification. Nevertheless, the encouraging outcomes achieved thus far indicate the viability of species-level classification using our approach with dual-polarization radar data.

Table 4
The results of experiments conducted on the dry beans dataset using the STNet, FNet, and BasicFNet models are presented. The best-performing model is bolded and highlighted in yellow, while the top-performing baseline is italicized and also highlighted in yellow.

	Model	Accuracy	Precision	Recall	F1
STNet	Baseline	0.15	0.02	0.14	0.04
	Co-learning	0.92	0.94	0.93	0.93
	Final Model	**0.93**	**0.94**	**0.94**	**0.94**
FNet	*Baseline*	*0.25*	*0.04*	*0.14*	*0.06*
	Co-learning	0.92	0.94	0.93	0.93
	Final Model	0.92	0.94	0.93	0.94
BasicFNet	Baseline	0.11	0.14	0.15	0.04
	Co-learning	0.92	0.93	0.93	0.93
	Final Model	0.92	0.94	0.93	0.94

Based on the Dry Beans Dataset: Table 4 shows the results of experiments performed using the Dry Beans dataset, which is the smallest in size among the other datasets. Consequently, the baseline experiment conducted on this dataset tends to exhibit significant overfitting, leading to test set results that are primarily random guessing, albeit with FNet showing a slight improvement, 25%. The introduction of additional labe1led data through the co-learning process once again enhanced the accuracy of these models. Moreover, the Final Models not only maintained the accuracy achieved through co-learning but also demonstrated a marginal improvement. This observation highlights the robust performance of the models, as a majority of the previously unlabelled data, affected by noise, has now been appropriately labelled.

Based on the Skin Segmentation Dataset: Table 5 shows the results of experiments performed using the Skin Segmentation dataset. Once again, the Baseline showed indications of overfitting on all three ensemble networks, primarily due to the relatively small labelled data. In this context, FNet demonstrated superior overall accuracies of 76%, followed by STNet and BasicFNet. This suggests that BasicFNet faces more challenges in precisely capturing and distinguishing the intricate features of the dataset, attributed to its smaller number of layers compared to FNet. The incorporation of co-learning strategies effectively mitigates overfitting in the baseline and enhances accuracy across all models. This highlights the significant efficacy of collaborative learning in addressing the initial challenges posed by noisy labelled data. The Final models achieve perfect accuracy through the co-learning process. This indicates that the co-learning approach effectively generalizes to an exceptional level of accuracy on the Skin Segmentation Dataset, particularly when more noisy data is labelled.

In this study, only 20% of the data was correctly labelled initially, meaning a relatively small portion of each dataset (especially the dry

Table 5
The results of experiments conducted on the skin segmentation dataset using the STNet, FNet, and BasicFNet models are presented. The best-performing model is bolded and highlighted in yellow, while the top-performing baseline is italicized and also highlighted in yellow.

	Model	Accuracy	Precision	Recall	F1
STNet	Baseline	0.50	0.63	0.65	0.45
	Co-learning	0.99	0.99	0.99	1.00
	Final Model	**1.00**	**1.00**	**1.00**	**1.00**
FNet	*Baseline*	*0.76*	*0.39*	*0.49*	*0.43*
	Co-learning	0.99	0.99	1.00	0.99
	Final Model	1.00	1.00	1.00	1.00
BasicFNet	Baseline	0.29	0.61	0.55	0.27
	Co-learning	0.99	0.99	0.99	0.99
	Final Model	1.00	1.00	1.00	1.00

beans dataset) had accurate labels. This limitation affected the Baseline model's ability to generalize effectively, resulting in lower initial accuracy. However, as more data was labelled during the co-learning process, accuracy improved, demonstrating the intended effectiveness of our models.

5.1. Detecting bird ring roosts and segmenting skin regions in pascal faces

We applied the STNet radar model to segment Purple Martin ring roosts, visible on the radar as doughnut-shaped structures due to dense bird aggregations, typically observed in late summer and early fall. Data from NEXRAD stations KHTX, KMHX, and KLVX were used and selected for their proximity to known bird activities. Radar products including reflectivity, differential reflectivity, and correlation coefficient were used to support visualization of predicted radar echoes, which are classified as weather (blue), other biology (purple), bats (brown), birds (yellow), and insects (green). The last column of Fig. 5 (highlighted in red) shows pixel-wise predictions for six radar elements, with examples of these products demonstrating the model's effectiveness. The STNet model detected the ring roosts as yellow doughnut shapes in the prediction masks, aligning with previously observed bird aggregation patterns. However, some misclassification were observed, with birds occasionally identified as other biological echoes, and weather signals containing bird-like signatures. These errors are reflected in the confusion matrix, which shows a small percentage of birds being misclassified as weather and vice versa. Despite these misclassification, the model's performance is promising, and could benefit from improvements that differentiates between bird and weather signals for more accurate detection in future radar-based studies.

We applied the STNet model, which demonstrated the best performance on the skin segmentation dataset, to segment images from the PASCAL FACE dataset introduced by Yan et al. [42]. This dataset, designed for face detection and recognition, is a subset of the PASCAL VOC dataset. The STNet model was trained using RGB pixel values, so it relies heavily on colour information for segmentation. The results show strong performance in accurately segmenting skin pixels based on colour. However, a key limitation is that the model sometimes misidentifies pixels with colours similar to human skin tones, belonging to non-skin objects. This is evident in the image in the last column, the second row of Fig. 6, where a wooden post in the background is incorrectly labelled as skin. These errors highlight a challenge in using RGB-based models for segmentation when the background contains colours that resemble skin tones. To reduce such misclassification, further refinements – such as incorporating additional features like texture or depth information – could improve segmentation accuracy.

5.2. Evaluation with state-of-the-art

In this section, we qualitatively evaluate our method in comparison to other state-of-the-art approaches, focusing on several key aspects critical to the performance of noise-handling methods.

5.2.1. Noise handling and robustness to noise

State-of-the-art noise-handling techniques address challenges in label correction, noise types, and noise rate tolerance. Label correction methods [18,20] use label noise modelling or transition matrices for precision in class-dependent scenarios. Noise-aware loss functions [23,30] robustly manage instance-independent and asymmetric noise, while adaptive strategies [28,43] excel in real-world asymmetric and instance-dependent noise. While most prior methods addressed label flip noise [18], symmetric noise [24,27], or complex instance-dependent noise [43], they primarily focused on moderate noise levels [22], but advanced techniques [28,43] handled extreme noise rates of up to 80% well. Our semi-supervised co-training method combines bootstrap categorical cross-entropy loss and ensemble models to achieve robustness with moderate-noise rates of 50% for domain-specific biological scatterer contexts. Most methods with direct applications to weather radar have typically adopted CNN methods. To effectively distinguish migration signals from background noise, [44] employed a semi-supervised learning approach, while [45] enhanced noise robustness by distinguishing biological echoes from meteorological noise using a superpixel-based method. [46] filters out ground clutter and precipitation noise using a 2.4° elevation angle and manual annotation, leveraging a random forest classifier for robustness. In comparison, our approach considers a broader range of scattered types in weather radar data than [44–46], while directly using radar data in its natural format without conversion to images. We only perform the conversion for result visualization.

5.2.2. Scalability across dataset sizes and complexities

Scalability across dataset sizes and complexities is crucial for noise-handling methods. Early approaches like [18] scale well to small and large benchmarks, while [22,23] show robustness on diverse, real-world datasets but struggle with complex domains. Enhancements by [25,27] improved performance across synthetic and real-world data. For large-scale datasets, [24,28] excel, while [30,43] improve performance on small- and large-scale datasets, with [43] achieving competitive results on challenging benchmark datasets. Our method extends scalability to domain-specific applications, effectively handling radar datasets labe1led with citizen-science data. This addresses domain-specific challenges where earlier methods did not attempt to address. Compared to other radar data methods, [44] labelled and processed large volumes of CINRAD weather radar data, generating additional images for three classes: birds, insects and precipitation. [45] compiled a large-scale dataset from 108 weather radar stations over two years, totalling 750,000 historical scans. [46] processed 4142 radar images, extracting 10 million insects and 6 million bird signatures. While most of these approaches relied on large labelled CINRAD datasets, our method, based on NEXRAD data, used only a few data points and focused on non-image data, significantly reducing storage and memory requirements.

5.2.3. Learning paradigms and adaptability in noise-handling techniques

Noise-handling techniques vary in adaptability and robustness. Early supervised methods like [18,22] use noise adaptation and loss correction but struggle with complex noise. [23] simplifies this with direct loss adjustments, while [20] improves accuracy via transition matrices. Dynamic methods such as [27] update labels during training, removing the need for clean data. Advanced strategies, like [24,30], combine supervised losses with selective learning for better generalization. Semi-supervised approaches [25,28] enhance adaptability by distinguishing clean from noisy data, while [43] uses adaptive label smoothing to dynamically adjust to noise. Our method builds on these by integrating semi-supervised co-training with a bootstrap ensemble, addressing domain-specific radar data challenges. While early methods establish foundational techniques, hybrid and semi-supervised approaches, including ours, offer superior robustness for real-world noisy datasets. In comparison, methods using radar data [44–46]

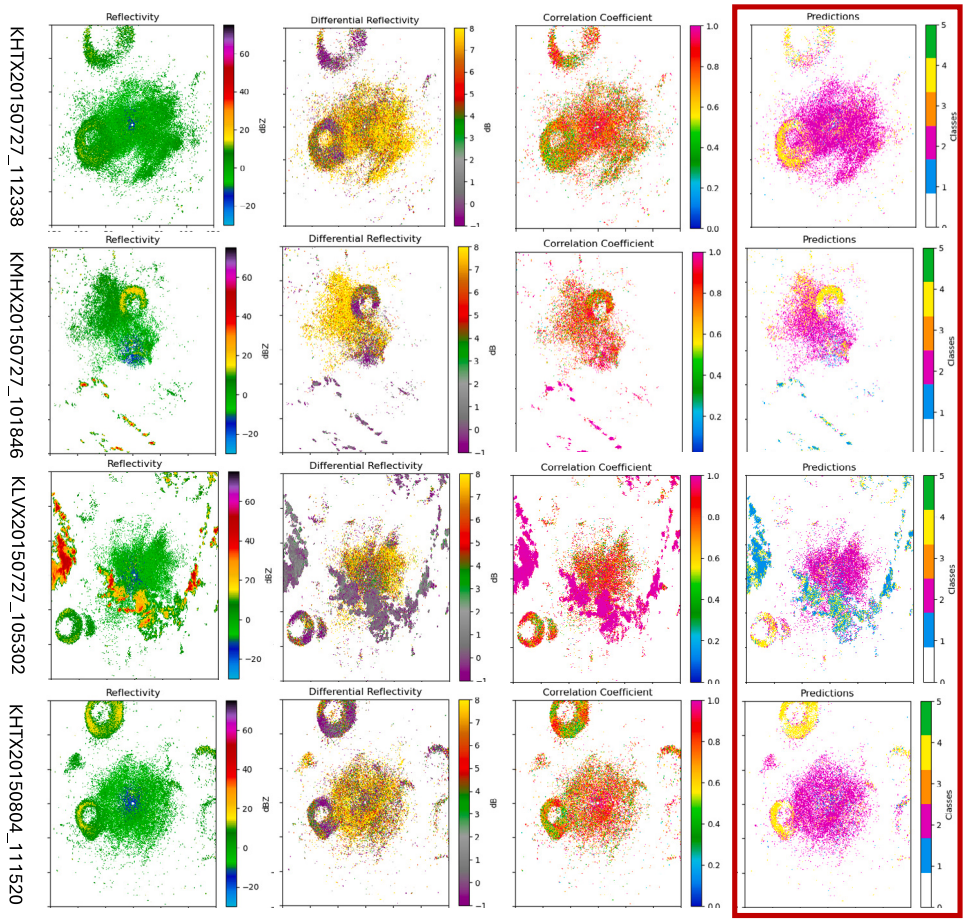

Fig. 5. The STNet radar model's pixel-wise predictions are shown in the last column, highlighted in red. To the left, samples from radar products—reflectivity, differential reflectivity, and correlation coefficient—depict radar echoes. White indicates areas without data, while echoes are classified as blue for weather, purple for other biology, brown for bats, yellow for birds, and green for insects. The yellow doughnut-shaped regions in the prediction mask represent successfully detected bird ring roosts. The labels to the left show the filename, which includes the radar name, date, and time of the scan.

Fig. 6. Example results of skin segmentation using the STNet model on sample images from the PASCAL FACE dataset [42]. The original images are displayed above their corresponding predicted segmentation masks. Skin pixels are shown in white, while the background is displayed in black. The segmented faces are highlighted with a red box.

employed a semi-supervised learning approach [44,45], utilizing both labelled and unlabelled data to improve adaptability in noise handling. Additionally, [46] used a supervised random forest classifier to mitigate overfitting caused by labelling noise. Among these, [44] placed a stronger emphasis on noise reduction, while the other closely related studies focused less on noise reduction compared to methods using different datasets.

5.2.4. Computational efficiency and model complexity

Methods for handling noisy data vary in computational efficiency and complexity, with many emphasizing minimal overhead and robust performance. Lightweight approaches like [18,27], and [20] use noise adaptation, matrix factorization, and simple adjustments, making them ideal for resource-limited scenarios. [23,30] focus on efficiency through lightweight loss functions, while [24,28] maintain low complexity, with [28] using semi-supervised learning, and [43] further employing adaptive label smoothing. Our method balances efficiency and performance by employing low-complexity ensembles and co-training, optimized for iterative refinement in domain-specific tasks like radar data. Methods like [22,25] avoid structural network changes, relying on loss function optimization, but approaches such as ours and [28] achieve a better trade-off with simple yet effective architectural adaptations. While earlier techniques excel in efficiency, our approach combines minimal complexity with domain-specific optimizations, excelling on medium-sized datasets. Compared to other methods using radar data, our approach is more computationally efficient as it avoids the high processing costs associated with CNNs. [44] employed CNN-based semi-supervised learning, converting radar products into six-channel RGB images, which significantly increased storage and computational complexity. In contrast, [45] used superpixel segmentation instead of CNNs, making it more efficient than [44]. However, our method further reduces computational demands by directly processing radar data without the need for image conversion.

6. Conclusion

In summary, our approach consistently demonstrated superior accuracy improvements across three datasets, affirming its efficacy in refining classification accuracy and rectifying noisy labels. Significantly, the STNet network ensemble consistently demonstrated superior performance, consistently achieving the highest accuracy in all conducted experiments. This consistent performance suggests the robustness of STNet in effectively addressing the challenges associated with noisy labels.

Our observations suggest that the overall accuracy improvement can be attributed to the bootstrapping ensemble co-training approach and the utilization of bootstrap sparse categorical cross-entropy loss. However, the distinctive performance of STNet is specifically attributed to the incorporation of temporal and frequency layers within its architecture. These layers significantly enhance the model's capacity to comprehend time-dependent and frequency-specific patterns in the data, enabling it to extract relevant features from the datasets.

Although the results obtained are promising, our study recognizes that there may be a potential for further improving the biological-level species echo classification. To address this, we suggest investigating the despeckling algorithm to mitigate certain misclassifications observed in our experiments, but this exploration falls beyond the scope of the current work. The despeckling algorithm, as successfully implemented in previous studies such as [4], holds promise for significantly improving the precision and reliability of our classification model.

In light of these findings, our results not only establish a strong foundation for the classification of biological species-level echoes using citizen science data but also indicate areas for future exploration and improvement. Our noise reduction approach sets the stage for ongoing studies aimed at advancing the methodology, thereby contributing to the evolving landscape of accurate and reliable classification techniques in machine learning for noisy label learning.

CRediT authorship contribution statement

John Atanbori: Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Investigation, Data curation, Conceptualization. **Christos A. Frantzidis:** Writing – review & editing, Writing – original draft. **Mohammed Al-Khafajiy:** Writing – review & editing, Writing – original draft. **Aliyu Aliyu:** Writing – review & editing, Writing – original draft. **Behnaz Sohani:** Writing – review & editing, Writing – original draft. **Kofi Appiah:** Writing – review & editing, Writing – original draft, Validation. **Harriet Moore:** Writing – review & editing, Writing – original draft. **Catherine Sanders:** Writing – review & editing, Writing – original draft. **Alastair I. Ward:** Writing – review & editing, Writing – original draft, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] J.D. Nichols, B.K. Williams, Monitoring for conservation, Trends Ecol. Evolut. 21 (12) (2006) 668–673.
[2] J.H. Verhagen, R.A. Fouchier, N. Lewis, Highly pathogenic avian influenza viruses at the wild–domestic bird interface in europe: Future directions for research and surveillance, Viruses 13 (2) (2021) 212.
[3] M.R. Kumjian, Principles and applications of dual-polarization weather radar. Part I: Description of the polarimetric radar variables., J. Oper. Meteorol. 1 (2013).
[4] A. Kilambi, F. Fabry, V. Meunier, A simple and effective method for separating meteorological from nonmeteorological targets using dual-polarization data, J. Atmos. Ocean. Technol. 35 (7) (2018) 1415–1424.
[5] S. Gauthreaux, R. Diehl, Discrimination of biological scatterers in polarimetric weather radar data: Opportunities and challenges, Remote. Sens. 12 (3) (2020) 545.
[6] P.M. Stepanian, K.G. Horton, V.M. Melnikov, D.S. Zrnić, S.A. Gauthreaux Jr., Dual-polarization radar products for biological applications, Ecosphere 7 (11) (2016) e01539.
[7] A.M. Dokter, P. Desmet, J.H. Spaaks, S. van Hoey, L. Veen, L. Verlinden, C. Nilsson, G. Haase, H. Leijnse, A. Farnsworth, et al., Biorad: biological analysis and visualization of weather radar data, Ecography 42 (5) (2019) 852–860.
[8] C. Chilson, K. Avery, A. McGovern, E. Bridge, D. Sheldon, J. Kelly, Automated detection of bird roosts using NEXRAD radar data and convolutional neural networks, Remote. Sens. Ecol. Conserv. 5 (1) (2019) 20–32.
[9] J. Meade, R. Van der Ree, P.M. Stepanian, D.A. Westcott, J.A. Welbergen, Using weather radar to monitor the number, timing and directions of flying-foxes emerging from their roosts, Sci. Rep. 9 (1) (2019) 10222.
[10] T.-Y. Lin, K. Winner, G. Bernstein, A. Mittal, A.M. Dokter, K.G. Horton, C. Nilsson, B.M. Van Doren, A. Farnsworth, F.A. La Sorte, et al., MistNet: Measuring historical bird migration in the US using archived weather radar data and convolutional neural networks, Methods Ecol. Evol. 10 (11) (2019) 1908–1922.
[11] K. Cui, C. Hu, R. Wang, Y. Sui, H. Mao, H. Li, Deep-learning-based extraction of the animal migration patterns from weather radar images, Sci. China Inf. Sci. 63 (2020) 1–10.
[12] S. Wang, C. Hu, K. Cui, R. Wang, H. Mao, D. Wu, Animal migration patterns extraction based on atrous-gated CNN deep learning model, Remote. Sens. 13 (24) (2021) 4998.
[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 248–255.
[14] N. Weisshaupt, T. Lehtiniemi, J. Koistinen, Combining citizen science and weather radar data to study large-scale bird movements, 2021.
[15] H. Song, M. Kim, D. Park, Y. Shin, J.-G. Lee, Learning from noisy labels with deep neural networks: A survey, IEEE Trans. Neural Netw. Learn. Syst. (2022).
[16] X. Liang, X. Liu, L. Yao, Review–a survey of learning from noisy labels, ECS Sensors Plus 1 (2) (2022) 021401.

[17] Z. Cheng, S. Gabriel, P. Bhambhani, D. Sheldon, S. Maji, A. Laughlin, D. Winkler, Detecting and tracking communal bird roosts in weather radar data, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 378–385.
[18] S. Sukhbaatar, R. Fergus, Learning from noisy labels with deep neural networks, arXiv preprint arXiv:1406.2080 2 (3) (2014) 4.
[19] Y. Wu, B. Li, Z. Li, Revising similarity relationship hashing for unsupervised cross-modal retrieval, Neurocomputing 614 (2025) 128844.
[20] Y. Yao, T. Liu, B. Han, M. Gong, J. Deng, G. Niu, M. Sugiyama, Dual t: Reducing estimation error for transition matrix in label-noise learning, Adv. Neural Inf. Process. Syst. 33 (2020) 7260–7271.
[21] D. Hendrycks, M. Mazeika, D. Wilson, K. Gimpel, Using trusted data to train deep networks on labels corrupted by severe noise, Adv. Neural Inf. Process. Syst. 31 (2018).
[22] G. Patrini, A. Rozza, A. Krishna Menon, R. Nock, L. Qu, Making deep neural networks robust to label noise: A loss correction approach, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1944–1952.
[23] Y. Xu, P. Cao, Y. Kong, Y. Wang, L_dmi: A novel information-theoretic loss function for training deep nets robust to label noise, Adv. Neural Inf. Process. Syst. 32 (2019).
[24] X. Ma, H. Huang, Y. Wang, S. Romano, S. Erfani, J. Bailey, Normalized loss functions for deep learning with noisy labels, in: International Conference on Machine Learning, PMLR, 2020, pp. 6543–6553.
[25] B. Han, G. Niu, X. Yu, Q. Yao, M. Xu, I. Tsang, M. Sugiyama, Sigua: Forgetting may make learning with noisy labels more robust, in: International Conference on Machine Learning, PMLR, 2020, pp. 4006–4016.
[26] J. Han, P. Luo, X. Wang, Deep self-learning from noisy labels, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 5138–5147.
[27] K. Yi, J. Wu, Probabilistic end-to-end noise correction for learning with noisy labels, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 7017–7025.
[28] S. Liu, J. Niles-Weed, N. Razavian, C. Fernandez-Granda, Early-learning regularization prevents memorization of noisy labels, Adv. Neural Inf. Process. Syst. 33 (2020) 20331–20342.
[29] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, M. Sugiyama, Co-teaching: Robust training of deep neural networks with extremely noisy labels, Adv. Neural Inf. Process. Syst. 31 (2018).
[30] D. Mandal, S. Bharadwaj, S. Biswas, A novel self-supervised re-labeling approach for training with noisy labels, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020, pp. 1381–1390.
[31] H. Wei, L. Feng, X. Chen, B. An, Combating noisy labels by agreement: A joint training method with co-regularization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 13726–13735.
[32] M.S. Van Den Broeke, Polarimetric radar observations of biological scatterers in hurricanes irene (2011) and sandy (2012), J. Atmos. Ocean. Technol. 30 (12) (2013) 2754–2767.
[33] J. Lee-Thorp, J. Ainslie, I. Eckstein, S. Ontanon, Fnet: Mixing tokens with fourier transforms, 2021, arXiv preprint arXiv:2105.03824.
[34] S. Reed, H. Lee, D. Anguelov, C. Szegedy, D. Erhan, A. Rabinovich, Training deep neural networks on noisy labels with bootstrapping, 2014, arXiv preprint arXiv:1412.6596.
[35] M. Koklu, I.A. Ozkan, Multiclass classification of dry beans using computer vision and machine learning techniques, Comput. Electron. Agric. 174 (2020) 105507.
[36] R.B. Bhatt, G. Sharma, A. Dhall, S. Chaudhury, Efficient skin region segmentation using low complexity fuzzy decision tree model, in: 2009 Annual IEEE India Conference, IEEE, 2009, pp. 1–4.
[37] P. Jatau, V. Melnikov, T.-Y. Yu, Detecting birds and insects in the atmosphere using machine learning on NEXRAD radar echoes, Environ. Sci. Proc. 8 (1) (2021) 48.
[38] J.D. Ray, P. Stepanian, J. Kelly, Evaluation of NEXRAD Radar as a Tool for Monitoring Monarch Butterflies, Tech. rep., Pantex Plant (PTX), Amarillo, TX (United States), 2019.
[39] P.M. Stepanian, S.A. Entrekin, C.E. Wainwright, D. Mirkovic, J.L. Tank, J.F. Kelly, Declines in an abundant aquatic insect, the burrowing mayfly, across major North American waterways, Proc. Natl. Acad. Sci. 117 (6) (2020) 2987–2992.
[40] P.B. Chilson, W.F. Frick, J.F. Kelly, K.W. Howard, R.P. Larkin, R.H. Diehl, J.K. Westbrook, T.A. Kelly, T.H. Kunz, Partly cloudy with a chance of migration: weather, radars, and aeroecology, Bull. Am. Meteorol. Soc. 93 (5) (2012) 669–686.
[41] W.F. Frick, P.M. Stepanian, J.F. Kelly, K.W. Howard, C.M. Kuster, T.H. Kunz, P.B. Chilson, Climate and weather impact timing of emergence of bats, 2012.
[42] J. Yan, X. Zhang, Z. Lei, S.Z. Li, Face detection by structural models, Image Vis. Comput. 32 (10) (2014) 790–799.
[43] Y. Li, H. Han, S. Shan, X. Chen, Disc: Learning from noisy labels via dynamic instance-specific selection and correction, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 24070–24079.
[44] Z. Sun, C. Hu, K. Cui, R. Wang, M. Ding, Z. Yan, D. Wu, Extracting bird and insect migration echoes from single-polarization weather radar data using semi-supervised learning, IEEE Trans. Geosci. Remote Sens. (2024).
[45] C. Hu, Z. Yan, K. Cui, R. Wang, J. Zhang, Z. Sun, D. Wu, Superpixel-based weak biological feature echo extraction method for weather radar, IEEE Trans. Geosci. Remote Sens. (2024).
[46] C. Hu, Z. Sun, K. Cui, H. Mao, R. Wang, X. Kou, D. Wu, F. Xia, Classification of biological scatters using polarimetric weather radar, IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. (2024).

John Atanbori received a Ph.D. in Computer Science from the University of Lincoln, UK, where he is currently a Senior Lecturer. He was previously a Computer Science lecturer at the University of Hull, UK. His research focuses on computer vision, machine learning, and deep learning. Before his lecturing career, he worked in the industry, developing computer vision and machine learning algorithms. He has also worked as a research fellow at the University of Nottingham's computer vision laboratory.

Christos A. Frantzidis is a Senior Lecturer in the School of Engineering and Physical Sciences at the University of Lincoln, UK, holds an MEng in Electrical and Computer Engineering from AUTH, Greece, an M.Sc. in Medical Informatics and a Ph.D. in Ageing Neuroscience from the AUTH Medical School. He has served as a guest scientist at the German Aerospace Agency on a European Space Agency-funded study and his latest research activities involve artificial intelligence applications in digital health, rehabilitation through artificial gravity and multi-modal language mapping (citations: 2936, h-index: 26, i10-index: 34).

Mohammed Al-khafajiy is a Senior Lecturer in Computer Science at the University of Lincoln and a member of the Lincoln Centre for Autonomous Systems (L-CAS). He has built a substantial portfolio of research in Cloud/Edge computing, Cognitive computing, Big Data mining, and embedded systems. Mohammed also actively contributes to international program committees and serves as a regular reviewer for esteemed journals such as IEEE Internet of Things (Q1), Future Generation Computer Systems (Q1), and Applied Sciences (Q2).

Aliyu Aliyu, is an Associate Professor in Sustainable Energy at the University of Lincoln, UK. His research integrates artificial intelligence and machine learning into energy systems for enhancing efficiency and sustainability. Dr Aliyu has published extensively in these areas and holds editorial roles in Frontiers in Chemical Engineering: Sustainable Process Engineering and the Nigerian Journal of Technological Development. He is interested in the digital transformation of legacy systems towards the energy transition.

Behnaz Sohani received her Ph.D. in Robotics and Biomedical Engineering from London South Bank University in 2020. She is currently a Lecturer in Robotics and Automation at Loughborough University and a member of the Intelligent Automation Centre. Previously, she was a Lecturer at the University of Lincoln, where she also completed a Postdoctoral Fellowship. Her research focuses on applying computer vision, machine learning, and AI to assistive and healthcare systems and robotics, including medical and surgical devices. Dr. Sohani is a Chartered Engineer and a Fellow of the Higher Education Academy and an active member of IEEE and IET.

Kofi Appiah is a Senior Lecturer at the University of York and studied for his Ph.D. in Computer Science at University of Lincoln. Before joining the University of York, Kofi worked at Sheffield Hallam University and Nottingham Trent University as a Lecturer/Senior Lecturer from 2013. He has previously worked with the Embedded and Intelligent Systems (ESI) Research Group at University of Essex in Colchester as Senior Research Officer and on part-time basis as a Development Engineer with Metrarc Ltd, Cambridge. Kofi has been a member of the IEEE since 2004.

Harriet Moore completed a Ph.D. in Science at the University of Melbourne and lectured in Human Geography of Lincoln until recently. She is currently a Senior Researcher in Geospatial Health and Well-being in the Lincoln Institute of Rural and Coastal Health. Her current research includes pioneering novel data science approaches for analysing ambulance data to identify communities and regions vulnerable to severe illness. Previously, she has analysed national river restoration datasets and worked closely with river management authorities in Australia snd the UK exploring economic, psychological, and institutional barriers to the effectiveness of environmental interventions, including the impact of drought on farmer well-being and willingness to maintain environmental projects.

Catherine Sanders received a Ph.D. in river science from Loughborough University, UK, and is currently a lecturer in physical geography and geography Department Lead at the University of Lincoln. Catherine was previously lecturer in physical geography and Director of Teaching and Learning at Canterbury Christ Church University. Catherine's research uses data driven approaches for conservation, with a focus on upscaling the understanding of animal–environment relationships for informing rewilding practices, community-driven conservation, and management of species invasions.

Alastair I. Ward received a Ph.D. in Environment from the University of York, UK in 2001. He is currently an Associate Professor at the University of Leeds and has previously been Head of Department at the University of Hull, and Head of Wildlife Research at the Animal and Plant Health Agency. His research into applied wildlife biology includes development of novel methods of wildlife surveillance to support disease modelling and decision-making.