RoBINN: Robust Bird Species Identification using Neural Network

Chirag Samal[1], Prince Yadav[2], Sakshi Singh[1], Satyanarayana Vollala[2] and Amrita Mishra[3]
[1]*Dept. of Electronics and Communication Engineering, International Institute of Information Technology, Naya Raipur, India*
[2]*Dept. of Computer Science and Engineering, International Institute of Information Technology, Naya Raipur, India*
[3]*International Institute of Information Technology, Bangalore, India*

Keywords: Deep Learning, Bird Species Identification, Speech Recognition, Convolutional Neural Network.

Abstract: Recent developments in machine and deep learning have made it possible to expand the realms of traditional audio pattern recognition to real-time and practical applications. This work proposes a novel framework for robust bird species identification using the neural network (RoBINN) based on their unique vocal signatures. To make the network robust and efficient, data augmentation is performed to create synthetic training samples for bird species with less available recordings. Further, inherent properties of audio signals are suitably leveraged via effective speech recognition-based feature engineering techniques to develop an end-to-end convolutional neural network (CNN). Additionally, the proposed model architecture for the CNN framework employs residual learning and attention mechanism to generate attention-aware features, which enhances the overall accuracy of birdcall identification. The proposed architecture employs an exhaustive dataset with 21375 recordings corresponding to 264 bird species. Experimental results validate the proposed bird species classification technique in terms of accuracy, F1-score, and binary cross-entropy loss.

1 INTRODUCTION AND RELATED WORKS

With the advancements in machine and deep learning (DL) techniques, audio and speech recognition have evoked widespread interest from industry and academia. To this end, bird species identification via their unique chirping sounds is one of the emerging applications of ML and DL in speech recognition. The number and diversity of bird species in an ecosystem is a critical indicator of the biodiversity and sustainability of the natural habitat (Priyadarshani et al., 2018). Thus, bird species identification has become fundamental to worldwide research related to the ecosystem's overall health and well-being.

Identifying bird species via appearance is tedious in nature and has led to the development of various audio-based recognition methods (Lasseck, 2018), (Sankupellay and Konovalov, 2018), (Chakraborty et al., 2016). However, classification via such traditional approaches are challenging due to the presence of background noise and multiple bird species in the same recording clip. Moreover, conventional audio recognition approaches that perform audio tagging, emotion and music classification, and sound event detection, etc. employ smaller datasets, thereby resulting in low accuracy (Kong et al., 2020).

In this regard, with the easy access of large datasets to the research community combined with the success of ML and DL in multi-faceted engineering paradigms, several recent works such as (Gyires-Tóth and Czeba, 2016) ,(Sang et al., 2018), (Xie et al., 2019), (Sankupellay and Konovalov, 2018) employ deep convolutional neural networks (CNNs) for bird species identification. However, the authors therein consider the classification of 11 and 46 species, respectively. Further, the results are observed to be less accurate with an increase in the number of species, thereby making their performances highly dependent on the number of bird species.

The work in (Zhao et al., 2017) employs the Gaussian mixture model along with support vector machines (SVMs) to perform the classification of 11 species. However, the proposed approach was observed to be inaccurate for multiple concurrent bird acoustic events. Next, the work (Chakraborty et al., 2016) performed classification of 26 species using dynamic kernel-based SVM and deep neural network (DNN). However, the authors therein did not account for the noisy environment condition. The subsequent

Samal, C., Yadav, P., Singh, S., Vollala, S. and Mishra, A.
RoBINN: Robust Bird Species Identification using Neural Network.
DOI: 10.5220/0010647500310038
In *Proceedings of the 18th International Conference on Signal Processing and Multimedia Applications (SIGMAP 2021)*, pages 31-38
ISBN: 978-989-758-525-8

work in (Kong et al., 2020) employed pre-trained audio neural networks (PANNs) on a very large scale audio dataset, and its architecture included Wavegram-Logmel-CNN. However, the approach achieved an accuracy of 43.9% on AudioSet tagging.

Thus, motivated by the success of DNNs, this paper develops a novel deep learning-based identification system for bird species in order to address the voids of the recent state-of-the-art works. Further, it is demonstrated that the proposed framework achieves desirable accuracy under noisy environments and can differentiate between the different species in a multi-species audio clip. The novelty and main contributions of the paper can be summarized as follows.

- A novel end-to-end bio-acoustic signal recognition system is developed for bird species classification using audio recordings containing multi-species birdcalls as well as accounting for a noisy environment.
- The robust bird species identification using neural network (RoBINN) architecture is developed wherein the attention mechanism is embedded into the network via residual learning to generate attention aware features and refine adaptive features.
- Exhaustive experiments demonstrate that the proposed RoBINN framework is precise as well as superior to state-of-the-art bird species identification architectures.

The usage of the term 'robust' in the proposed design has a two-fold significance. Firstly, RoBINN can successfully perform bird species identification in challenging noisy environments. Secondly, by virtue of the residual learning and attention mechanism aspects in the proposed architecture, complex spectral features present in the audio clips can be identified to yield more precise results.

The remainder of the paper is organized as follows. Section II describes the overall workflow of the proposed RoBINN framework for bird species identification followed by the detailed model architecture in Section III. The experimental results and discussion are presented in Section IV followed by the concluding remarks in Section V.

2 OVERALL WORKFLOW OF THE PROPOSED ROBINN FRAMEWORK

This section presents a brief overview of the various components involved in the proposed deep learning-based bird identification framework.

2.1 Dataset Description

The dataset used in this work is 'Cornell Bird Challenge' (CBC)' dataset (Cornell Lab of Ornithology, 2020) obtained from xeno-canto.org (Planqué et al., 2005). The dataset includes 264 unique bird species with 21375 recordings belonging to different bird sounds. The training data consists of both audio data and metadata which includes various parameters such as type of audio channel, length, altitude, dates, location, etc. The audio data is sampled at 8 different sampling frequencies, out of which more than 95% are sampled at 44.1 kHz and 48 kHz. Further, visual information of about 75.8% of the birds is available while recording the training data at 64 different locations, each associated with the latitude and longitude of the concerned area. The duration of each audio signal in the training data varies from 5 seconds to 500 seconds.

2.2 Data Preprocessing

The training data comprises of audio files in 8 different sampling frequencies which are re-sampled to a common sampling frequency of 32 kHz. Subsequently, short-time Fourier transform (STFT) is employed to obtain the spectrogram of the entire audio event in a windowed manner. The STFT parameters such as the window width and degree of overlap between successive windows are chosen similar to (Zhao et al., 2017).

2.3 Data Augmentation

For training a DNN, a large amount of data samples are required corresponding to each of the bird species. The training dataset comprises of some bird species with fewer samples. Thus, data augmentation approach has been employed to create new and synthetic training samples by adding small perturbations to the initial training set. The various data augmentation techniques employed in this work are discussed as follows.

2.3.1 Standardization

Standardization technique generates a synthetic sample $\bar{x}$ as

$$\bar{x} = \frac{x - \mu}{\sigma} \tag{1}$$

where x, μ, and σ denote the actual sample point, mean and standard deviation corresponding to all the samples in the dataset.

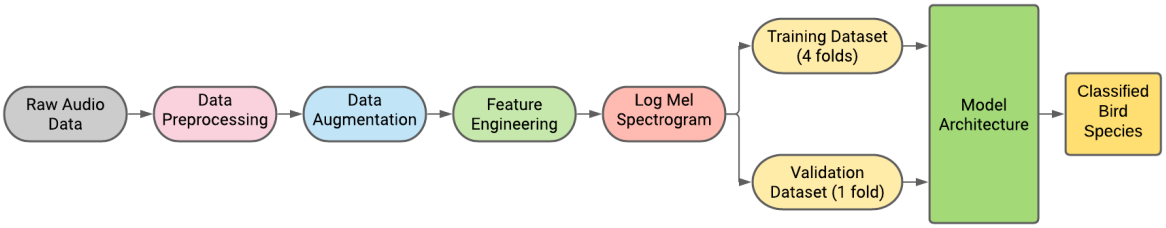


Figure 1: Work flow diagram for proposed DL-based bird identification framework: RoBINN.

2.3.2 Adding White Noise

To enhance the performance of the proposed model under noisy conditions, additive white Gaussian noise (AWGN) is added to the audio signal data.

2.3.3 Audio Shifting

The audio clip is shifted towards the left or right direction with a random time scale. For example, if the audio is shifted towards left or right with x seconds, the first or the last x seconds will be marked as zero, respectively. This work considers a random shifting of the bird sounds by 5000 μs similar to the related work in (Schlüter and Grill, 2015).

2.3.4 Audio Stretching

The speed of the audio clip is modified by a factor without affecting the pitch. A factor of 1.2 is considered for the work under discussion similar to the related work in (Schlüter and Grill, 2015).

2.3.5 Changing the Pitch

The pitch of the sound is either lowered or raised by keeping the speed of sound constant. Each sample was pitch shifted by 10 values ranging from -10 to 10.

2.3.6 Mixup

Mixup corresponds to training of the DNN on convex combinations of pairs of training examples and their corresponding classes (Zhang et al., 2017). The virtual training sample $\bar{x}$ is constructed by employing the input vectors x_i and x_j as

$$\bar{x} = \lambda x_i + (1-\lambda)x_j \tag{2}$$

where $\lambda \in [0,1]$ denotes a tuning parameter and its corresponding label y can be obtained as

$$\bar{y} = \lambda y_i + (1-\lambda)y_j \tag{3}$$

where y_i and y_j denote one-hot label encodings.

2.3.7 SpecAugment

Similar to the work in (Park et al., 2019), using time and frequency masking techniques, this data augmentation technique is applied to the log-mel spectrogram of an audio clip.

2.4 Feature Engineering

Feature engineering, one of the most essential components of an ML process, creates features that enable the simpler functioning of the ML-based algorithms by using relevant domain knowledge. Thus, the performance of any ML algorithm critically depends on the features on which the training and testing are performed.

The basic spectral properties are obtained by transforming the time-based raw audio signal into its frequency domain counterpart via the Fourier transform. A frequency representation is essential for extraction of useful characteristics of the signal. This signal is further divided into segments to get a spectrum, which determines the strength of the signal over various frequencies present in the waveform. Further, the frequencies present in the segments are converted into log-mel scales. One of the major steps is to pass the spectrum through the mel-frequency filters/analyser to get the log-mel spectrums (Kahl et al., 2019) as it concentrates on only certain spectral components. As a function of frequency regions, these filters are spaced non-uniformly on the frequency axis. Log-mel spectrograms are visual representations of frequency differences in the received log-mel spectrums.

2.5 Log-mel Spectrogram

The mel-scaled spectrogram is evaluated employing the time-series input and subsequently this power spectrogram is converted to decibel (dB) units. The minimum and maximum frequencies for the threshold are set as 50 Hz and 16 kHz respectively. The spectrogram is then resized to 224*224 to fit into the proposed network architecture. The other parameters

selected for the generation of the mel spectrograms are sampling rate of 32 KHz and 64 number of mel bins. A window length of 1024 is chosen which describes the size of the window employed to evaluate discrete Fourier Transform (DFT) of the audio signal. The hop length which refers to the extent of the window shift along the audio signal during the processing of the short-term Fourier transform is set as 320. Let the mel-spectrum be represented as $X[k]$. The logarithm of the mel-spectrum $\log X[k]$ can be expressed as (Yan et al., 2019)

$$\log X[k] = \log H[k] + \log E[k] \quad (4)$$

where $H[k]$ and $E[k]$ denote the spectrum envelope and spectrum details, respectively. On taking Inverse DFT of the above equation one obtains,

$$x[k] = h[k] + e[k], \quad (5)$$

where $x[k]$ is referred to as cepstrum, $h[k]$ represents the spectral envelope and is termed as the Mel-frequency cepstral coefficients (MFCC) .
A block diagram representation of the overall workflow of the proposed DL-based bird identification framework: RoBINN is given in Fig. 1.

3 SYSTEM OVERVIEW

This section details the various components of the proposed model architecture.

3.1 Mish Activation Function

Conventional ResNet-based architectures typically employ rectified linear unit (ReLU) activation function. However, a major disadvantage associated with the ReLU activation function is that it loses the gradient information caused by the breakdown of negative inputs to zero. To overcome this shortcoming, several activation functions have been proposed in literature namely Leaky ReLU, ELU, SELU, swish, Mish etc. Despite Mish having the most complex interference, it has been demonstrated to provide better empirical results in comparison to others (Misra, 2019). Thus, this work employs the Mish activation function in the basic block. Mish is a non-monotonically self-regulating activation function and is unlimited in the positive direction i.e can have much higher positive values (Misra, 2019). The slight allowance for negative values enables the network for better gradient flow in comparison to absolute zero limit as in ReLU. For an input x, the mish function $f(x)$ can be mathematically defined as [give ref]

$$f(x) = x * \tanh(\ln(1 + e^x)). \quad (6)$$

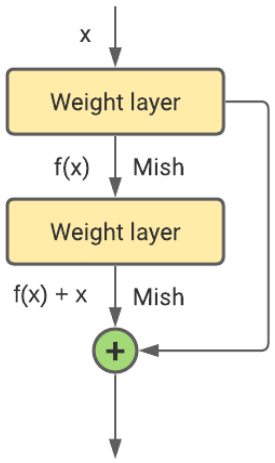


Figure 2: Basic block for residual learning.

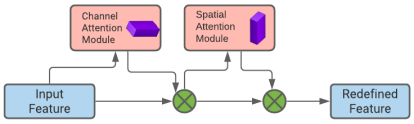


Figure 3: General architecture of convolutional block attention module (CBAM).

3.2 Residual Blocks

Residual blocks are skip-connection blocks used in the input layer to learn residual functions. The output mapping $H(x)$ can be expressed as $f(x) + x$, where x denotes the residual. The basic idea behind residual block is that optimising residual mapping is simpler than optimising unreferenced mapping. Fig. 2 gives a pictorial representation of the basic residual learning block wherein the activation function from the previous layers is being added to the activation of a deeper layer in the network. This work employs the Mish activation function in both basic and residual blocks.

3.3 Convolutional Block Attention Module (CBAM)

It's an attention module for the CNN framework (Woo et al., 2018). CBAM converts intermediate function maps into spatial and channel dimensions to derive attention maps. The attention map is then element-wise multiplied with the input function map to produce a refined and highlighted output. Let the intermediate feature map be represented as $F \in R^{C \times H \times W}$, where C, H, and W denote the number of channels, height, and width of the feature map, respectively.

The channel and spatial attention maps can be expressed as $M_c \in R^{C \times 1 \times 1}$ and $M_s \in R^{1 \times H \times W}$, respec-

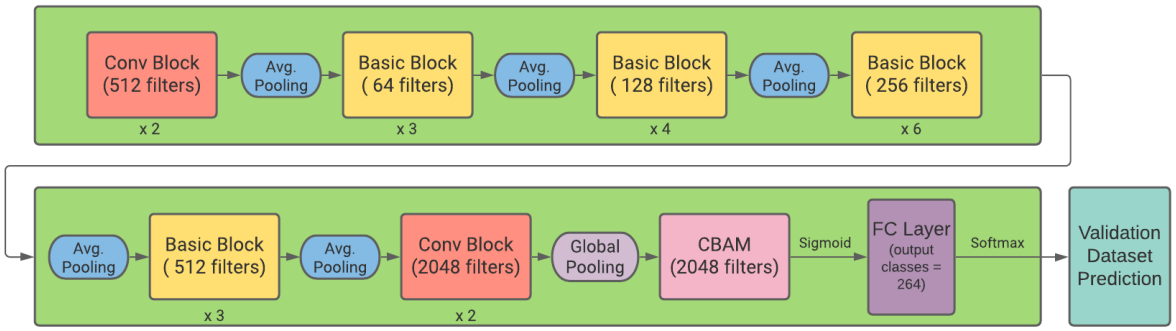


Figure 4: Model architecture of RoBINN.

tively. The entire process can be expressed as

$$F' = M_c(F) \otimes F \tag{7}$$

$$F'' = M_s(F') \otimes F' \tag{8}$$

where F' and F'' denote channel-refined feature, final refined output respectively and $\otimes$ represents element wise multiplication. The attention values are scattered along the spatial dimension during multiplication. The channel and spatial modules can be arranged in a sequential or parallel manner. Research shows that sequential arrangement tends to achieve better result in comparison to the parallel one. Further, in sequential arrangement, channel first-order is marginally better in comparison to spatial first-order.

3.4 Model Architecture

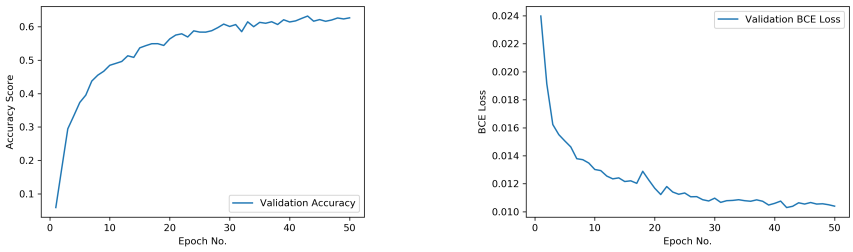
This section describes the proposed model architecture in detail. Fig. 4 depicts the block diagram representation of the architecture. It consists of a total number of 38 layers embedded with an attention module. Log-mel spectrogram with 1000 frames and 64 mel bins of the audio data is taken as input in the designed model architecture. The proposed model architecture is parameterized as follows:

- l_1-l_2: Each of the two layers is represented by a ConvBlock which comprises of a convolution layer with kernel size 3×3 and 512 channel filters. The output representation of the convolution layer is subsampled using the average pooling of size 2×2.
- l_3-l_8: Represented by three sequential basic blocks which is a skip-connection block that consists of two convolutional layers of 64 channel filters and 3×3 kernel size.
- l_9-l_{16}: Represented by 4 sequential basic blocks with 128 channel filters of kernel size 3×3.
- l_{17}-l_{28}: Represented by 6 sequential basic blocks. Each convolutional layer consists of 256 channel filters of kernel size 3×3.
- l_{29}-l_{34}: 3 sequential basic blocks are used to represent these layers with 512 channel filters of kernel size 3×3.
- l_{35}-l_{36} : Represented by two ConvBlocks. The output feature map from l_{36} is subsampled via global average pooling to generate a fixed length vector from the feature maps.
- *CBAM* : To generate attention aware features, it sequentially infers attention maps along two different dimensions, channel and spatial, from an input function map. It is followed by the sigmoid function.
- l_{37}-l_{38} : In this, a fully connected layer of 2048 hidden units is added to extract the embedding features to enhance representation ability of the signal. This is followed by another fully connected layer which incorporates softmax function to calculate the probabilities for the bird species.

Here, each convolutional layer is followed by batch normalization (Ioffe and Szegedy, 2015) and then non-linear Mish activation function is employed. Dropout is applied after every downsampling operation and fully connected layer to prevent overfitting of the network.

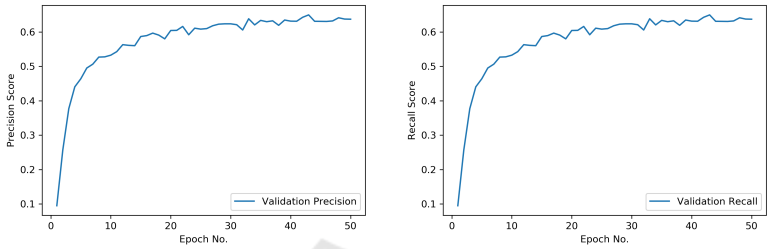
4 EXPERIMENTAL RESULTS AND DISCUSSION

In this section, after the initial training on the recordings of the train data, the experimental results for the validation set are obtained and analyzed. The dataset used in this study is the 'Cornell Bird Challenge' (CBC)' dataset (Cornell Lab of Ornithology, 2020), which was collected from xeno-canto.org (Planqué et al., 2005). The dataset is highly skewed since the number of recordings for a few species was very low, while others had a large number of recordings. Further, the recordings differed in length, quality and the amount of noise. The various parameters used for the

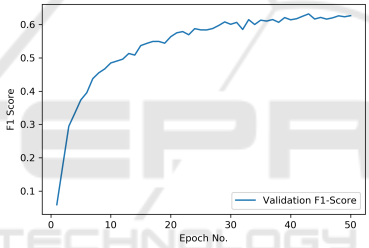


(a) (b)

Figure 5: (a) Accuracy score plot (b) Binary cross-entropy plot.



(a) (b)



(c)

Figure 6: (a) Precision score plot (b) Recall score plot (c) F1-score plot.

proposed RoBINN framework are summarized in Table 1.

Table 1: Parameters for experimental setup.

Model Configuration	
Parameter	**Value**
Sample Rate	32000 Hz
Window Size	1024
Hop Size	320
Mel-bins	64
Minimum Frequency	50 Hz
Maximum Frequency	14000 Hz
Number of Classes	264
Dropout Probability	0.4
Maximum Epochs	50
K-cross validation	5
Input Spectrogram Size	224
Learning Rate	0.001

The model is trained for 50 epochs using K-cross validation with $K = 5$. The binary cross-entropy (BCE) loss function is chosen for optimization of the weights associated with the proposed RoBINN framework. This particular choice of the loss function is made owing to the following reasons. Firstly, it is well-known that the BCE loss function is evaluated for every output class and is independent of the other classes. Secondly, for a multi-label classification problem, the NN should be designed such that the condition that an element belongs to a particular class does not influence the decision for other classes. Let an audio clip be denoted by x_n, where n is the index of the audio clip and $y_n \in \{0,1\}^k$ denotes the label of x_n. The quantity $f(x_n) \in [0,1]^k$ corresponds to the soft probabilities provided by the model where k represents the total number of classes. The BCE loss function l can be evaluated as (Yi-de et al., 2004)

$$l = -\sum_{n=1}^{N} (y_n \ln f(x_n) + (1 - y_n) \ln(1 - f(x_n))) \quad (9)$$

where N is the total number of audio clips.The Adam

Table 2: Test F1-score comparison on the Cornell BirdCall Challenge dataset.

Paper	Methodology	F1-score
(Incze et al., 2018)	A MobileNet pre-trained CNN model employing spectrograms of the audio clips.	0.474
(Knight et al., 2020)	An AlexNet-based architecture which employs spectrograms of different frequency and amplitude scales for bird species classification.	0.423
(Koh et al., 2019)	An inception-v3 model with Adam optimizer which enhanced the results of the baseline ResNet-18 architecture.	0.567
(Sankupellay and Konovalov, 2018)	The ResNet-50 architecture was used to automate the identification of bird species using audio syllables as the basic recognition unit.	0.614
Proposed RoBINN	An automated bird species identification framework based on residual learning and attention mechanism.	0.632

optimizer is employed instead of the conventional stochastic gradient descent to update the weights of the network during the training period. Hyperparameters like the learning rate and momentum are tuned during the training process to achieve accurate results. The initial learning rate is set as 0.001 and subsequently adjusted based on the loss function. Further, in order to facilitate identification of primary as well as any secondary bird species which may be present in the audio recording, RoBINN performs both frame and clip-level classification. Each audio clip is segregated into multiple frames. A classifier is employed per frame to yield the class existence probabilities followed by an aggregation of the classifier results for the entire audio clip.

The following performance metrics are employed for evaluation of the RoBINN framework:

- Classification Accuracy: Ratio of correctly classified bird species samples to the total number of samples.
- Precision : The ratio of true positives to the sum of true positives and false positives.
- Recall : The ratio of true positives to the sum of true positives and false negatives.
- F1-score : Harmonic mean of recall and precision.

Experiments were carried out on the training set and evaluated on the test set. The various performance metrics discussed earlier have been demonstrated for the test dataset in figures 5(a)-(b) and 6(a)-(c). The accuracy score versus number of epochs is demonstrated in Fig. 5(a). It can be observed that the accuracy score converges to a value of 0.64374 after 50 epochs.The BCE loss plot from Fig. 5(b) is observed to attain a value of 0.0103. In addition, the precision and recall scores after each epoch have been plotted in figures 6(a) and 6(b) respectively. Further, the precision and recall scores are observed to converge to 0.6435 and 0.6356, respectively after 50 epochs. The maximum F1-score achieved by the proposed RoBINN framework is 0.63221 as demonstrated in Fig.6(c).

Table 2 illustrates a comparison of the proposed RoBINN framework with state-of-the-art bird species identification systems. F1-score is employed to evaluate various models owing to the imbalanced class distribution of the bird species classification problem. In order to perform a fair comparison, only the proposed architecture is replaced with the respective neural networks of the existing works while the remaining parameters are kept same as that of RoBINN. It can be concluded from Table 2 that the proposed RoBINN framework exhibits a superior bird species identification in comparison to the existing studies.

5 CONCLUSIONS AND FUTURE SCOPE

Recent developments in the field of deep and machine learning have encouraged researchers to re-examine the traditional approaches to solve a wide range of engineering problems. Towards this end, bird species identification, a real-time application of audio/speech recognition has generated a lot of interest owing to its importance in studies related to the well-being of natural habitat. This paper proposes an end-to-end automatic bird identification technique RoBINN, which employs some aspects of conventional speech recognition via feature engineering into a robust and efficient deep neural network. Further, the model architecture incorporates residual learning and attention mechanism to generate attention aware features for enhancing the accuracy of bird species recognition. Exhaustive experimental results validate the optimality and precision of the proposed RoBINN framework

for bird species identification in comparison to the existing works. In conclusion, the proposed method can be deemed suitable for practical bio-acoustic monitoring of bird species in terrestrial environments and can be considered for deployment on a wider scale.

Future extensions of the work will explore validation of the proposed RoBINN framework using an external dataset and study of recent transformer-based DL model for bird species identification.

REFERENCES

Chakraborty, D., Mukker, P., Rajan, P., and Dileep, A. D. (2016). Bird call identification using dynamic kernel based support vector machines and deep neural networks. In *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 280–285. IEEE.

Cornell Lab of Ornithology (2020). Cornell Birdcall Identification. https://www.kaggle.com/c/birdsong-recognition. Online; accessed 15 October 2020.

Gyires-Tóth, B. and Czeba, B. (2016). Convolutional neural networks for large-scale bird song classification in noisy environment.

Incze, A., Jancsó, H.-B., Szilágyi, Z., Farkas, A., and Sulyok, C. (2018). Bird sound recognition using a convolutional neural network. In *2018 IEEE 16th International Symposium on Intelligent Systems and Informatics (SISY)*, pages 000295–000300. IEEE.

Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

Kahl, S., Stöter, F.-R., Goëau, H., Glotin, H., Planque, R., Vellinga, W.-P., and Joly, A. (2019). Overview of birdclef 2019: large-scale bird recognition in soundscapes. In *Working Notes of CLEF 2019-Conference and Labs of the Evaluation Forum*, number 2380, pages 1–9. CEUR.

Knight, E. C., Poo Hernandez, S., Bayne, E. M., Bulitko, V., and Tucker, B. V. (2020). Pre-processing spectrogram parameters improve the accuracy of bioacoustic classification using convolutional neural networks. *Bioacoustics*, 29(3):337–355.

Koh, C.-Y., Chang, J.-Y., Tai, C.-L., Huang, D.-Y., Hsieh, H.-H., and Liu, Y.-W. (2019). Bird sound classification using convolutional neural networks. In *CLEF (Working Notes)*.

Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., and Plumbley, M. D. (2020). Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894.

Lasseck, M. (2018). Audio-based Bird Species Identification with Deep Convolutional Neural Networks. In *CLEF (Working Notes)*.

Misra, D. (2019). Mish: A self regularized non-monotonic neural activation function. *arXiv preprint arXiv:1908.08681*.

Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., and Le, Q. V. (2019). Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*.

Planqué, B., Vellinga, W., Pieterse, S., and Jongsma, J. (2005). Xeno-canto: sharing bird sounds from around the world.

Priyadarshani, N., Marsland, S., and Castro, I. (2018). Automated birdsong recognition in complex acoustic environments: a review. *Journal of Avian Biology*, 49(5):jav–01447.

Sang, J., Park, S., and Lee, J. (2018). Convolutional recurrent neural networks for urban sound classification using raw waveforms. pages 2444–2448.

Sankupellay, M. and Konovalov, D. (2018). Bird call recognition using deep convolutional neural network resnet-50. In *Proc. ACOUSTICS*, volume 7, pages 1–8.

Schlüter, J. and Grill, T. (2015). Exploring data augmentation for improved singing voice detection with neural networks. In *ISMIR*, pages 121–126.

Woo, S., Park, J., Lee, J.-Y., and So Kweon, I. (2018). Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19.

Xie, J., Hu, K., Zhu, M., Yu, J., and Zhu, Q. (2019). Investigation of different cnn-based models for improved bird sound classification. *IEEE Access*, 7:175353–175361.

Yan, G., Wang, M., Liu, X., and Song, X. (2019). Sound event recognition based in feature combination with low snr. In *2019 International Conference on Artificial Intelligence and Advanced Manufacturing (AIAM)*, pages 109–114.

Yi-de, M., Qing, L., and Zhi-Bai, Q. (2004). Automated image segmentation using improved PCNN model based on cross-entropy. In *Proceedings of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing, 2004.*, pages 743–746. IEEE.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. (2017). mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

Zhao, Z., Zhang, S.-h., Xu, Z.-y., Bellisario, K., Dai, N.-h., Omrani, H., and Pijanowski, B. C. (2017). Automated bird acoustic event detection and robust species classification. *Ecological Informatics*, 39:99–108.