IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 23 September 2022, accepted 2 November 2022, date of publication 9 November 2022,
date of current version 15 November 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3220735

RESEARCH ARTICLE

Residual Learning for Marine Mammal Classification

DANIEL T. MURPHY[1], ELIAS IOUP[2], MD TAMJIDUL HOQUE[1], AND MAHDI ABDELGUERFI[1]
[1]Canizaro Livingston Gulf States Center for Environmental Informatics, The University of New Orleans, New Orleans, LA 70148, USA
[2]Center for Geospatial Sciences, Naval Research Laboratory, Stennis Space Center, MS 39529, USA

Corresponding author: Md Tamjidul Hoque (thoque@uno.edu)

This work was supported in part by the U.S. Navy (Office of Naval Research) under Contract N00173-16-2-C902.

ABSTRACT The passive acoustic monitoring of marine mammals is an essential tool for researchers tracking the populations of individual species in threatened environments. Given the large quantity of audio data generated by passive acoustic arrays, it is desirable to automate the process of identifying marine mammals present in the recordings. Utilizing acoustic data from the William A. Watkins Marine Mammal Sounds Database, we present an approach using residual learning networks (ResNets) for classifying the marine mammal vocalizations of up to 32 species. We first determine the optimal methods for converting acoustic recordings into discrete spectrograms suitable for input into neural networks. A series of configurations for spectrographic window functions, preprocessing augmentations, and multi-channel spectrogram generation are examined. Each configuration's spectrographic output is used to train a residual learning network. Its multi-class classification performance is ranked using the harmonic mean of precision and recall to calculate a weighted F1-score. Configurations specifying 512 $\times$ 256 spectrograms created with a Hann window of 1024 and utilizing horizontal roll demonstrate superior performance. We use the top-performing configurations to generate training data as input for a series of single and multi-channel residual neural networks. These networks are trained to high precision before evaluating their multi-class classification performance. A single-channel network performed the best, obtaining an F1-score of 0.867 with an AUC of 0.9281 on a 32-class classification task. Our multi-channel configuration obtained an F1-score of 0.846 with an AUC of 0.9169. While we demonstrate that networks may learn more information from multi-channel spectrographic inputs, we find that single-channel spectrograms offer superior classification performance overall.

INDEX TERMS Machine learning, marine mammal, vocalization, classification, ResNet, residual learning, multi-channel, spectrogram.

I. INTRODUCTION

Over 51% of marine mammal environments are threatened by climate change, pollution, by-catch, and other sources [1]. Conservationists working to mitigate these threats require up-to-date population data to make informed policy decisions [2]. Acoustic data containing marine mammal vocalizations may be used to effectively estimate marine mammal population changes [2], [3]. These changes, in turn, may be used as general indicators of overall ecosystem health [3]. Passive acoustic monitoring (PAM) is a popular tool for capturing this population data. By recording and identifying local marine mammal vocalizations, PAM is capable of accurately tracking both the presence and migration of these threatened species [4]. While PAM is an effective source of population data, it suffers from a problem of scale. The increase in the capacity and affordability of data storage parallels an increase in the number of PAM networks being deployed annually [4]. As a result, the amount of data currently being collected exceeds the capacity of researchers who possess the domain

The associate editor coordinating the review of this manuscript and approving it for publication was Binit Lukose.

knowledge necessary for making accurate classifications [5]. An automated solution is required.

A variety of computational solutions have been developed, with researchers demonstrating a steady improvement in classification performance over time [6]. Early machine-learning approaches were effective but required the manual selection of temporal, spectral, and or statistical features [6].

The introduction of Convolutional Neural Networks (CNNs) demonstrated improved classification performance while automating the feature selection process. Researchers implementing increasingly deeper CNNs initially saw steady gains in classification performance [7]. Unfortunately, these gains peaked or even diminished in sufficiently deep networks. He et al. highlighted this problem, demonstrating that a 56-layer CNN resulted in significantly higher training and test errors compared to a 20-layer CNN [8]. This decrease in performance is caused by the vanishing gradient problem whereby identity mappings learned in early layers are not propagated through the network [8]. For our model architecture, we chose a deep convolutional neural network (DCNN) that utilizes residual learning, in part for its ability to overcome the vanishing gradient issue. A residual learning network (ResNet) solves this problem by using residual (skip) connections to maintain early identity mappings through the network [8]. Additionally, ResNet has been proven to excel at spectrographic classification tasks [9] and shown to be an excellent candidate for marine mammal vocalization specifically [10], outperforming other DCNNs [11].

ResNets have been shown to outperform traditional CNNs at vocalization classification tasks generally [12], but they are also vulnerable to overfitting on small datasets [13], [14]. As such, we examine the vocalization performance of both ResNets and a traditional CNN. For this comparison, we needed to use an architecture known to be highly performant on large multi-class vocalization classification tasks. Accordingly, we based our model on a CNN designed by Sprengel et al. that won the 2016 BirdCLEF competition [15].

While CNNs were developed for the efficient analysis of images [8], they may be used for other graphical representations as well [16]. In the case of audio data, a discrete Fourier transform (DFT) is used to convert audio files into fixed-size spectrograms [17]. Several studies have been conducted utilizing ResNets to classify spectrograms of marine mammal vocalizations by species, but these studies are limited to identifying small (three or fewer) sets of species [18].

We endeavor to classify a larger range of species, which results in several novel challenges. Our data represent sets of marine mammal families with significantly different vocalizations. Classifying between such disparate vocalizations may seem trivial but distinguishing between individual species in those sets requires our ResNets to identify subtle timbral variations. Our networks demonstrate an ability to learn these subtleties across several sets of marine mammals simultaneously.

A secondary component of our research is the use of multi-channel spectrograms. Since ResNets may accept multi-channel inputs [8] and STFT spectrograms contain only one channel, we explore approaches for using additional channels effectively, ultimately combining them into a single, multi-channel spectrogram for input into ResNets. Our goal is to determine whether multiple channels allow a network to learn additional information, then to evaluate whether this increase in information allows for better classification performance than single channel representations. Previous studies on marine mammal vocalizations have utilized multi-channel spectrograms, but like studies on ResNets, they too were limited to three or fewer species [6].

In this paper, the following contributions are made.

1) We recreate a CNN design that has been proven to perform well on vocalization classification tasks. By comparing this top-performing CNN architecture to our ResNet, we demonstrate a significant improvement in multi-class classification performance using residual learning.
2) We investigate the potential utility of multi-channel spectrograms for classifying up to 32 distinct marine mammals. We ultimately find that while surplus information is gained from additional channels, this doesn't necessarily result in an overall improvement in performance compared to single-channel implementations.
3) We provide a brief outline of our findings for optimal spectrographic preprocessing configurations, with suggestions for optimal spectrogram size, window functions, and preprocessing techniques in the context of marine mammal audio data.

II. BACKGROUND

A. MARINE MAMMAL VOCALIZATIONS

Marine mammals utilize vocalizations for navigation, prey detection and avoidance, and communication [4]. Given the relatively low attenuation of acoustic energy in water, vocalizations provide a reliable method for detecting marine mammals over a wide area [19]. A machine learning model may be trained to classify these detections, but this requires a large dataset of vocalizations from the targeted species. For our research, we utilize audio recordings from the William A. Watkins Marine Mammal Sound Database. These recordings span seven decades and represent a variety of marine environments and recording technologies [20]. This variety allows our models to train on samples that contain differing levels of biophonic, geophonic, and anthrophonic noise.

Our audio samples represent vocalizations from two Orders of marine mammals, Cetaceans and Pinnipeds. Cetaceans are made up of two parvorders, Odontocetes (toothed whales, dolphins, and porpoises) and Mysticetes (baleen whales). Our samples also represent two Pinniped families, odobenids (walrus) and phocids (true seals). Together, these vocalizations exhibit considerable variation, with frequencies ranging from infrasonic Mysticete

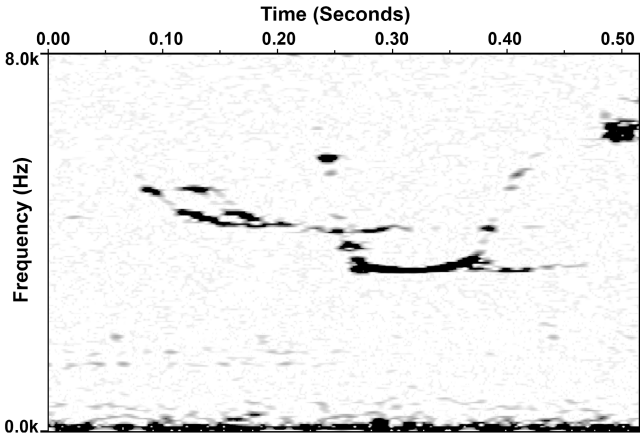

FIGURE 1. Whistle vocalizations from an atlantic spotted dolphin. Consisting of sustained tones with varying frequencies, whistles vary significantly from clicks.

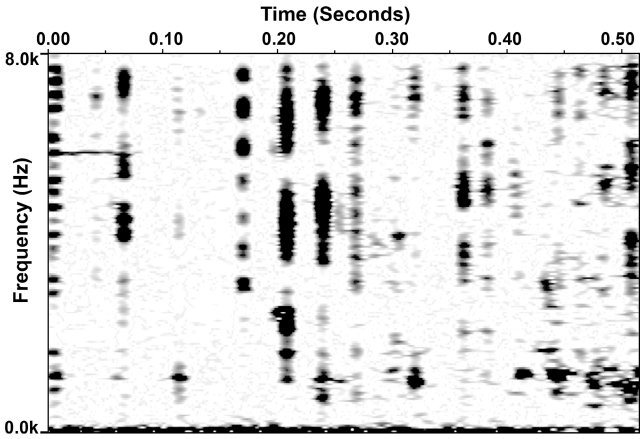

FIGURE 2. Click vocalizations from an atlantic spotted dolphin exhibiting short, broadband pulses.

calls reaching 17 Hz to odontocete pulsed clicks exceeding 25,000 Hz [4]. While considerable similarity may be found between closely related species [21], significant variation is seen even in vocalizations between the same species. The extent of same-species vocalization variation is seen in the markedly different spectrograms generated from Atlantic Spotted Dolphins' whistles FIGURE 1 and clicks FIGURE 2.

B. DIGITAL SIGNAL PROCESSING

CNNs may take spectrograms as input. In our study, recordings are imported at a sample rate of 22,050 Hz. These samples are then converted into spectrograms using a DFT (1).

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

For the k^{th} frequency where $X[k]$ represents the k^{th} Fourier coefficient.

As the magnitude of frequencies varies significantly across the duration of marine mammal vocalizations, it is necessary to perform multiple DFTs on sub-intervals of the original sample to capture the dynamic progression of frequency magnitude [17]. A window function is used to sample these sub-intervals. The window's width specifies the number of samples used to generate each spectral estimate. There are two trade-offs to consider when choosing the function type and width. 1. The size of the window represents a trade-off between spectral resolution and statistical variance. 2. The window function determines its shape, which in turn represents a trade-off between smearing and spectral leakage [22]. For example, the Bartlett window used in FIGURE 3 is a triangular function that results in less spectral leakage at the cost of smearing [23]. There is no best-choice for window widths and functions, rather, choices must be determined from the data being analyzed. For example, in FIGURE 3 we see that a narrow window may be necessary to effectively study rapid pulsed vocalizations, but the same window may be insufficient to discern the subtle frequency variations seen in the sustained tones of FIGURE 1. Multi-channel spectrograms provide a possible solution to this problem, allowing us to use different window functions for each channel.

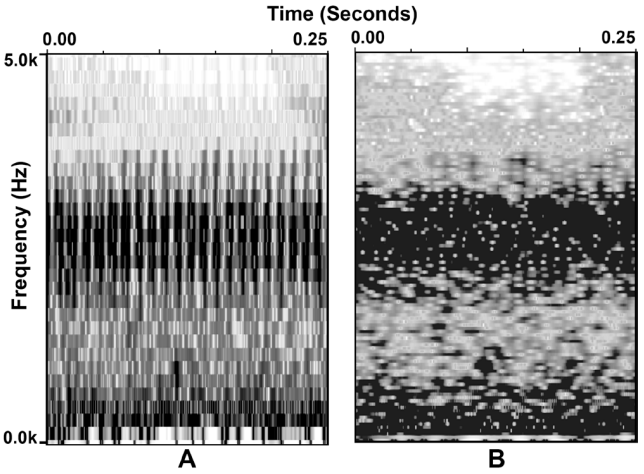

FIGURE 3. Spectrograms, created from the same rapid pulsed vocalization of an atlantic spotted dolphin, (A) uses a narrow band Bartlett window of size 256, allowing for the resolution of pulses, (B) uses a bartlett window of size 1024, making it difficult to resolve individual pulses.

C. MULTI-CHANNEL SPECTROGRAMS

Typically, discrete spectrographic data takes the form of a matrix containing the relative magnitude of frequencies across time. Multi-channel spectrograms add a channel-dimension. A multi-channel spectrogram may be created by stacking a series of single channel spectrograms across this channel dimension, resulting in a matrix that contains multiple representations of fixed length acoustic data in a single sample. Researchers have taken several approaches towards utilizing these channels.

In a marine mammal study, Thomas, Martin, Kowarski, et al. generated multiple channels by varying the window function used to generate each channel's spectrogram [18]. This approach provides a possible solution to the window function

resolution trade-off discussed in II.B. Thomas et al. found an increase in performance using this technique, but their study was limited to a 3-class classification task involving closely related species. We extend this approach to include up to 32 classes.

Another approach involves combining multiple audio sources of a single acoustic event. Researchers have demonstrated the utility of this approach in the analysis of data where spatiality is an important component, such as acoustic scene classification [24], [25]. As our sample data contains only monaural recordings, we don't explore this approach further.

Researchers studying human speech have demonstrated the utility of measuring the rate of change of frequency magnitude with respect to time [26]. These delta measures have been used in a variety of tasks, including dialect recognition, parts of speech classification, and even detecting emotional states [27], [28], [29]. There is a lack of studies investigating the utility of magnitude-delta measures for marine mammal vocalizations. As such, we include delta and delta-delta channels in our analysis of multi-channel spectrograms.

D. RESIDUAL LEARNING

While CNNs have been shown to be effective at classifying vocalizations [16], sufficiently deep implementations suffer from a degradation of training and test accuracy. This degradation is not due to overfitting and increases with the addition of more layers [30]. In their 2015 paper, He, Zhang, Ren, et al. demonstrated that this degradation was due to a loss of feature information mapped early in the networks [8]. They proposed a method to retain identity mappings while also training successively deeper layers via residual learning blocks (or ResNet blocks). These blocks may train two or more convolutional layers alongside a residual connection that maintains the identity of the original input [8]. This approach is defined formally in (2).

$$y = F(x, (W_i)) + x \tag{2}$$

The function $F(x,(W_i))$ represents the residual mapping to be learned, where x is the original input vector. Adding F(x) and x together is performed by element-wise addition, as seen in FIGURE 4.

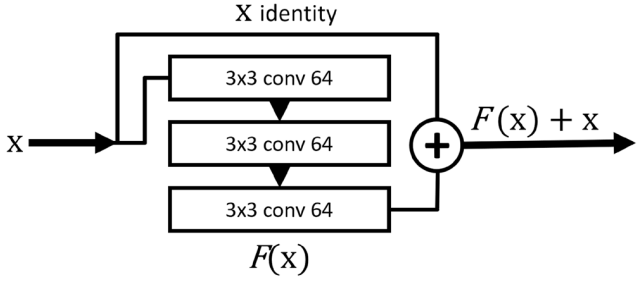

FIGURE 4. A representation of a 3-layer ResNet block with a residual connection preserving the original identity mapping.

E. EVALUATING MODEL PERFORMANCE

We utilize k-fold cross-validation to evaluate our models. In this process, we group our data into training and test sets k times and use these groups to train k different models. We evaluate the performance of each of these models and combine the results, providing us with a more reliable estimate of overall predictive performance [6].

To determine the performance of a model, we use a combination of two metrics: precision and recall. Precision measures the number of correct, positive predictions, while recall measures the number of positive samples that are correctly predicted. Since a model may improve precision at the expense of recall, and vice versa, it is necessary to consider both measures together to evaluate predictive performance effectively— this may be done by calculating the harmonic mean of precision and recall, a measure known as the F1-score. F1-score serves as the primary metric for our evaluations. Definitions for these metrics are given in Table 1.

To evaluate classification performance across all 32 species under study, an overall score may be calculated by taking each class's mean scores. As there is some class imbalance in our dataset, we weigh each score relative to its class size. Our method for calculating weighted precision is outlined in Table 1, but weighted measures for other scores are calculated similarly.

TABLE 1. Terms for multi-class classifier evaluation.

Term	Definition	Equation
Precision	Measures the accuracy of positive predictions	$\frac{TP}{TP + FP}$
Recall	Measures positive samples that are correctly predicted	$\frac{TP}{TP + FN}$
F1-score	Harmonic mean of precision and recall	$\frac{TP}{TP + \frac{1}{2}(FP + FN)}$
Weighted Precision	Measures precision of all classes with respect to their relative size. For i classes of size w_i with precision p_i	$\frac{\sum_{i=1}^{n} w_i p_i}{\sum_{i=1}^{n} w_i}$

TP = True Positive, TN = True Negative, FP = False Positive, FN False Negative

III. RELATED WORKS

The classification of marine mammal vocalizations using machine learning techniques is the subject of active research, with several recent studies suggesting novel optimization algorithms [31], model selection techniques [32], and spectrographic representations [18] worthy of further study.

The effectiveness of ResNet based models in positively identifying cetacean calls is demonstrated in *ORCA-SPOT: An Automatic Killer Whale Sound Detection Toolkit Using Deep Learning:* Bergler et al. [11]. The authors utilize a large dataset of Orca recordings to train a variety of ResNet architectures. Interestingly, they found that while ResNet18

was highly accurate, achieving an accuracy of 95.48% on a binary classification task, deeper implementations improved accuracy by less than 0.5% while taking significantly longer to train.

Buchanan et al. [10] demonstrate that ResNet-18 outperforms other DCNNs on the task of detecting bottlenose dolphins. They present compelling evidence for choosing ResNet architectures for vocalization classification tasks, but this study too is limited to the detection of a single species vocalizing in the high sonic to ultrasonic (3,000Hz - 144,000Hz) range exclusively. Further study on a wider frequency range is desirable.

A study by Thomas et al. [18] addresses the effectiveness of ResNets in classifying marine mammals with lower frequency vocalizations: blue whale, sei whale, and fin whale. This study is of particular interest for our research because the authors not only demonstrate the effectiveness of ResNets in classifying vocalizations but utilize a novel multi-channel spectrographic representation as well. They find that multi-channel spectrograms offer a modest improvement over single-channel, however, their results are limited to 3 species, each with distinct vocalization types. These papers make a compelling case for the use of multi-channel ResNets. We hope to further these findings by testing classification performance for a more diverse array of marine mammals exhibiting a larger range of vocalization types and frequencies. Additionally, we want to know if such a model will also be able to distinguish between species with closely related vocalization types, for example, pulsed clicks in a similar frequency range.

IV. METHODS

A. DATA

Previous studies utilizing residual learning for classifying marine mammal audio have focused on a limited number of classes, typically three or less [18], [33]. As such, we are interested in exploring classification performance for a broader range of species. The William A. Watkins Collection of Marine Mammal Sound Recordings provides an ideal source of labeled data for this task. We utilize data from the "best of" section of their sound database which contains recordings of relatively higher quality and lower noise and represents 32 species identified with high confidence [20]. The species include members of the Odontocete and Mysticete suborders of the infraorder Cetacea as well as the Phocid and Otariid families of the clade Pinnipedia. The recordings span seven decades and thus comprise a variety of recording technologies, ambient noise levels, and sample rates. The audio was recorded and annotated by William Watkins, William Schevill, G. C. Ray, D. Wartzok, D. and M. Caldwell, K. Norris, and T. Poulte and is freely available for academic use [20], [34].

B. PREPROCESSING DATA

Our models are trained on normalized spectrographic data stored as a series of 1-channel or 3-channel spectrograms of fixed width and height. We utilize two types of channels: spectrographic channels and delta channels. These channels may be combined to yield a 3-channel spectrogram. Spectrographic channels utilize a DFT to convert the original lossless audio file into a spectrogram. Each spectrographic channel may have its own parameters specified, including hop length, the width of the windowed signal before and after padding, and the window function. Once the spectrogram has been created, its amplitude values are converted into decibels using the function: $10 \cdot log(x)$.

Delta channels represent the rate of change across time of a previously generated spectrographic channel. Delta values are calculated by taking the change in magnitude with respect to time over an incrementing, fixed interval, as shown in Algorithm 1 below. Once the initial 3-channel spectrogram is created, further augmentations may be made to the data by the processes outlined later in this section.

Algorithm 1 Magnitude-Delta Calculation

Input: spectrogram
ns = new spectrogram of same dimensions
For i from 0 to (spectrogram.width – interval)
  For e from 0 to spectrogram.height
    $ns_{e,(i+interval)} = (s_{e,(i+interval)} - s_{e,i})^2$
Return ns normalized

A three-step process performs noise removal. A mask is generated by setting all magnitude values above a threshold to 1 and setting those below to 0. Gaussian smoothing is applied to the mask. Looping samples is occasionally necessary when the original recording is too short a length to fill the desired spectrogram width. A loop function duplicates the spectrogram across the time axis when this occurs until no gap remains.

A vertical roll function may be applied to samples that are at least 1.5 times as wide as the desired spectrogram width. Given a desired width w, a spectrogram s and an interval i such that $i = w/2$, n sub-samples of s are created spanning the following indices:

$$(s_{1i}, s_{1i+w}), (s_{2i}, s_{2i+w}), \ldots (s_{ni}, s_{ni+w})$$

Since some audio recordings in the Watkins Marine Database may be significantly longer than the database average, we limit the number of sub-samples taken to at most 5, to avoid excessive class imbalance and computational complexity.

Once augmentations (if any) have been applied, each channel's data is normalized. The channels are combined into a sample consisting of a 3-channel spectrographic matrix and a class label. Finally, all the completed samples are exported as NumPy (.npy) files, and all relevant configuration details are logged.

C. CONVOLUTIONAL NEURAL NETWORK

To establish a performance baseline to measure our ResNets against, we created a CNN based on the top-performing entry

in the 2016 BirdCLEF competition [15]. It takes as input a 128 × 256 spectrogram and consists of 5 convolutional layers followed by a fully connected layer and a final fully connected layer which uses a SoftMax function to output the predicted class. A summary of its structure is seen in FIGURE 5.

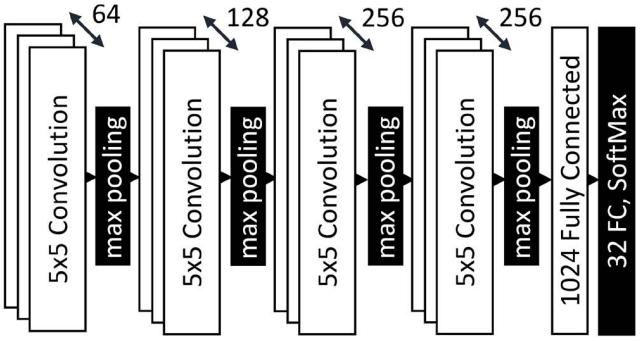

FIGURE 5. CNN implementation consists of 4 sets of 5 × 5 Convolutions.

D. RESIDUAL LEARNING NETWORK

ResNets have been shown to improve classification performance over traditional CNNs [8]. This improvement stems from the use of residual connections that maintain identity mappings across blocks of traditional convolutional layers, as seen in Figure 7. The benefits of residual learning extend to marine mammal vocalization studies. However, while researchers have shown ResNets to perform well on marine mammal classification for small (three or fewer) sets of closely related marine mammal species [18], a study for a larger range of species is needed. Given the success of ResNets in large avian classification tasks [35], residual learning is a promising candidate for large-scale marine mammal classification.

Our residual learning network (ResNet) implementations are based on the paper published by He, et al., the creators of ResNet [8]. Our ResNets are made up of three consecutive sets of one or more residual blocks each. For each consecutive set, filter size (height and width) decreases while filter depth increases, as seen in FIGURE 6. These sets are followed by a fully connected layer and a final fully connected layer using the SoftMax function to output the predicted class.

Each residual block contains a 3 × 3 convolutional layer with batch normalization and a Relu activation function. This is followed by another 3 × 3 convolutional layer whose output is batch normalized and summed with the identity mapping output of the previous block. This sum is input into a Relu activation function, and finally, the output is sent to the next residual block. This residual block structure is shown below in FIGURE 7. Dimensionality reduction between stacks is handled by a pooling layer, implemented as a convolutional layer with a stride of 1 in the first residual block of each stack.

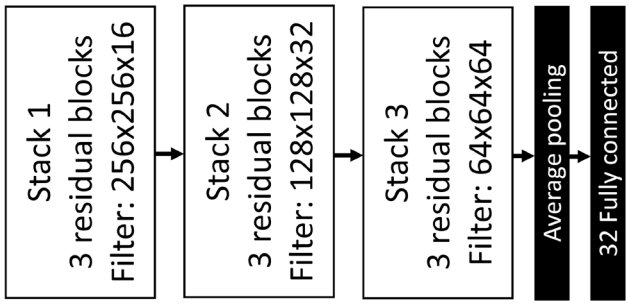

FIGURE 6. 3-channel residual learning network implementation.

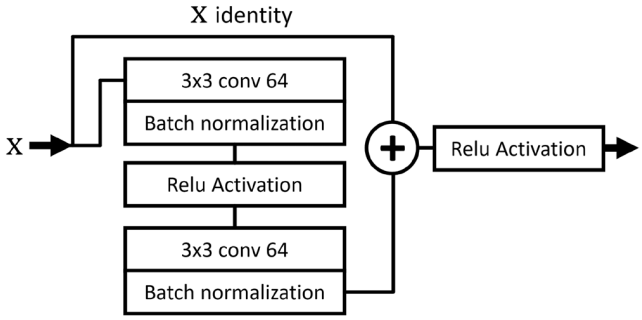

FIGURE 7. Design of a single residual block containing 2 convolutional layers.

V. RESULTS AND DISCUSSION

This section contains the results of our analysis of a CNN, and our 1-Channel and 3-Channel ResNets on the task of classifying marine mammal species via vocalizations. This is followed by a deeper analysis of optimal spectrographic parameters for 1 and 3-channel ResNets, and a discussion of our findings. Measures of accuracy are calculated as the mean of k-folds. Precision, recall, and f1-scores are calculated by taking the mean of k-folds from the mean of all classes, weighted by their respective class size. Values for confusion matrices are summed from each k-fold. Unless otherwise noted, we use 5-fold cross-validation across 100 epochs on a 16-class classification task as the standard for comparing optimal parameters. Our final test in section IV-E utilizes 10-fold cross-validation across 250 epochs on a 32-class classification task.

A. CONVOLUTIONAL NEURAL NETWORK

As a baseline, we conduct a comparison between a top-performing CNN architecture and a 1-Channel ResNet. As seen in TABLE 2, ResNet shows a significant improvement over CNN across all measures.

TABLE 2. Comparison of CNN and ResNet.

	f1-Score	Precision	recall	accuracy
CNN	0.8019	0.8179	0.8045	0.8045
1ch. ResNet	0.8512	0.8599	0.8543	0.8543

B. MULTI-CHANNEL INFORMATION GAIN

To determine if ResNets may learn more information when using multiple spectrographic channels, we conducted a comparison between a test and control spectrogram. Our test spectrogram contains three channels, each generated from a unique configuration. We evaluated each of these channel configurations on a single-channel classification task. Of these three, we selected the best performing configuration and used it to generate a control spectrogram with three identical channels. In Table 3, we see that the test spectrogram does demonstrate superior performance, indicating a positive gain in information.

TABLE 3. Results for test and control multi-channel spectrographic configurations.

	f1-score	precision	recall
Test	0.7090	0.7274	0.7101
Control	0.5951	0.6571	0.6010

C. OPTIMAL PREPROCESSING PARAMETERS

Results for our analysis of optimal spectrographic preprocessing parameters are outlined below. These findings serve as a basis for the design of our final 1 and 3-channel ResNet implementations.

In TABLE 4 we see that the top-performing window function was a Hann window of size 1024 with a hop length of 128. Note that no window function outperformed the others across all parameter combinations.

TABLE 5 shows results for optimal spectrogram width and height, with superior performance found using size 512×256. These spectrograms were generated using a Hann window function with a window length of 1024 and a hop length of 64.

The horizontal role provides superior results, as seen in TABLE 6. Note, in this test, only one roll is performed.

D. SINGLE AND MULTI-CHANNEL RESNET RESULTS

TABLE 7 contains the classification performance results for a series of 1 and 3-channel configurations. A wide variety of window and channel configurations are tested to reveal any trends, such as an increase in window size, that may impact classification performance. These results, together with our analysis of optimal preprocessing parameters, inform the configurations used in our final, 32-class classification tests in section V.E.

E. 32-CLASS CLASSIFICATION RESULTS

The above results informed the creation of three final spectrographic configurations. Their performance on a 32-class marine mammal vocalization classification task is seen in TABLE 8. We trained these models for 300 epochs using 10-fold cross-validation. While we rely upon F1-scores to evaluate performance, we include area under the curve (AUC) scores for reference in these final tests. Training and validation curves are provided in Figure 8.

The results seen in TABLE 8 indicate that single-channel configurations offer superior performance, with the

TABLE 4. Prediction scores for top-performing window functions.

f1-score	window	window-length	hop-length	spec. dimensions
0.90047	hann	1024	128	512×256
0.89666	ham	1024	128	512×256
0.88532	triangle	1024	128	512×256
0.87063	triangle	1024	64	512×256
0.86873	ham	1024	64	512×256
0.86351	hann	1024	64	512×256
0.83379	triangle	256	64	512×256
0.83261	hann	256	64	512×256
0.8317	hann	512	32	512×256
0.83068	ham	512	32	512×256
0.8282	ham	256	64	512×256
0.81822	triangle	1024	32	512×256
0.81282	ham	1024	32	512×256
0.81049	hann	1024	32	512×256
0.80483	ham	128	64	512×256
0.80235	triangle	512	32	512×256
0.78723	triangle	128	64	512×256
0.78662	hann	128	64	512×256
0.75283	hann	128	32	512×256
0.74721	triangle	128	32	512×256
0.74042	triangle	1024	16	512×256
0.73628	ham	128	32	512×256
0.72828	ham	1024	16	512×256
0.7021	hann	1024	16	512×256

TABLE 5. Spectrographic dimension results.

f1-score	height	width
0.86617	512	256
0.86497	128	256
0.85518	256	256
0.83623	512	128
0.82666	64	512
0.82336	32	1024
0.82115	64	1024
0.79275	128	128
0.77059	256	512
0.64117	64	64

TABLE 6. Horizontal roll results.

h-roll	f1-score	precision	recall
yes	0.8968	0.9018	0.8975
no	0.8178	0.8398	0.8199

TABLE 7. Results for optimized single and multi-channel ResNet configurations.

f1-score	channel type(s)	window-length(s)	hop-length	dimensions
0.87995	[w]	512	64	256×256
0.85721	[w][w][w]	512, 256, 128	64	256×256
0.84146	[w]	128	256	64×64
0.83236	[w]	256	64	128×128
0.82304	[w][d][d]	512	64	256×256
0.81801	[w][d][w]	512, n/a, 128	64	256×256
0.80912	[w][d][d]	256	64	128×128
0.80627	[w][w][w]	128, 64, 32	256	64×64
0.7913	[w][w][w]	256, 128, 64	64	128×128
0.79024	[w][d][w]	128, n/a, 32	256	64×64
0.7811	[w][d][w]	256, n/a, 64	64	128×128
0.74639	[w][d][d]	128	256	64×64

[w] = window channel, [d] = delta channel

TABLE 8. Results for 32-class classification task.

f1-score	auc	channel	h-roll	window-length(s)	hop-length	dims.
0.86715	0.9281	[w]	yes	1024	64	512×256
0.85496	0.9286	[w]	yes	1024	64	128×256
0.84558	0.9169	[w][d][d]	yes	1024	64	128×256

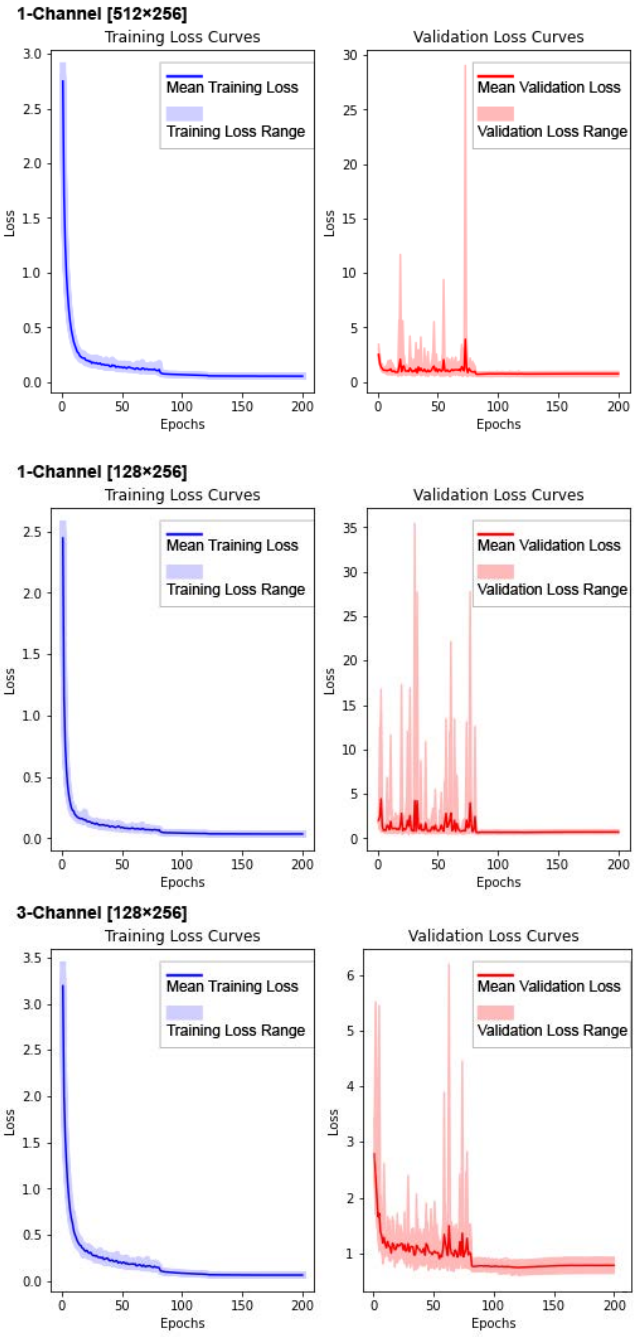

FIGURE 8. Training and validation loss curves for the final 3 models on a 32-class classification task. The lighter areas represent the total range of the loss curves across ten folds. The solid lines indicate the mean.

512 × 256 spectrogram implementation demonstrating a modestly higher f1-score than the others. These results may be surprising given the positive information gain demonstrated in section V.B. One possibility is that, given the relatively low sample rate (22,050 Hz) in our data, the increase in temporal resolution gained from using separate window functions provides only limited utility. A better use case may be found in studies using data with higher sample-rates. Given a broader range of frequencies, varying the window function across channels may have more of an impact on learning. In particular, data containing ultrasonic vocalizations could see a benefit. Further, multi-channel spectrograms may be a good choice for classifying between small numbers of closely related species, as seen in previous studies [18].

Our results do indicate useful techniques for improving classifier performance: augmenting data with the horizontal roll and increasing frequency resolution both demonstrate advantages. Our implementation of horizontal roll results in functionally more samples being trained, so this result is not surprising. The impact of frequency resolution does offer some interesting insights, however. The best performing spectrographic window functions utilized the highest frequency resolutions available given their relative height. This suggests that subtle variations in frequency are a large factor in distinguishing between species. We can further conclude that the diminished temporal resolution of the top-performing configuration was still sufficient to resolve a

significant number of rapid pulses and clicks contained in the recordings.

VI. CONCLUSION

Both traditional CNNs and ResNets are effective at classifying marine mammal vocalizations, however, ResNets demonstrate superior results on our dataset. Multi-channel spectrograms demonstrated an increase in overall information gain compared to a control, however, multi-channel implementations failed to outperform single-channel implementations on a 32-class classification task. More research is necessary, but given the results for our dataset, we conclude that the added complexity of multi-channel spectrographic input is not justified for marine mammal species classification using acoustic data.

We find that several preprocessing factors lead to improved classification results. Increasing frequency resolution was associated with an increase in performance across all our tests, with the best results being given by Hann window functions of a size of 1024. Increasing the size of the spectrograms was generally associated with an increase in performance. It should be noted that our tests were limited by memory constraints to the maximum tested sizes of 512×256 or 256×512. It is likely that performance would continue to improve with larger spectrograms of higher resolutions, but further research is needed to confirm this trend. Finally, horizontal-roll improved classification with a single horizontal roll giving a 9.7% increase in F1-score.

REFERENCES

[1] I. C. Avila, K. Kaschner, and C. F. Dormann, "Current global risks to marine mammals: Taking stock of the threats," *Biol. Conservation*, vol. 221, pp. 44–58, May 2018.

[2] B. Hendricks, E. M. Keen, J. L. Wray, H. M. Alidina, L. McWhinnie, H. Meuter, C. R. Picard, and T. A. Gulliver, "Automated monitoring and analysis of marine mammal vocalizations in coastal habitats," in *Proc. OCEANS*, Kobe, Japan, May 2018.

[3] L. Hatch, C. Wahle, J. Gedamke, J. Harrison, B. Laws, S. Moore, J. Stadler, and S. Van Parijs, "Can you hear me here? Managing acoustic habitat in U.S. waters," *Endangered Species Res.*, vol. 30, pp. 171–186, May 2016.

[4] W. W. Au and M. C. Hastings, *Principles of Marine Bioacoustics*. New York, NY, USA: Springer, 2008.

[5] K. D. Seger, M. H. Al-Badrawi, J. L. Miksis-Olds, N. J. Kirsch, and A. P. Lyons, "An empirical mode decomposition-based detection and classification approach for marine mammal vocal signals," *J. Acoust. Soc. Amer.*, vol. 144, no. 6, pp. 3181–3190, Dec. 2018.

[6] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY, USA: Springer, 2009.

[7] A. Géron, *Hands-on Machine Learning With Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. Sebastopol, CA, USA: O'Reilly Media, Inc, 2019.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[9] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "CNN architectures for large-scale audio classification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, New Orleans, LA, USA, Mar. 2017, pp. 131–135.

[10] C. Buchanan, Y. Bi, B. Xue, R. Vennell, S. Childerhouse, M. K. Pine, D. Briscoe, and M. Zhang, "Deep convolutional neural networks for detecting dolphin echolocation clicks," in *Proc. 36th Int. Conf. Image Vis. Comput. New Zealand (IVCNZ)*, Dec. 2021, pp. 1–6.

[11] C. Bergler, H. Schröter, R. X. Cheng, V. Barth, M. Weber, E. Nöth, H. Hofer, and A. Maier, "ORCA-SPOT: An automatic killer whale sound detection toolkit using deep learning," *Sci. Rep.*, vol. 9, no. 1, pp. 1–17, Dec. 2019.

[12] C.-Y. Koh, J.-Y. Chang, C.-L. Tai, D.-Y. Huang, H.-H. Hsieh, and Y.-W. Liu, "Bird sound classification using convolutional neural networks," in *Proc. CLEF (Working Notes)*, Lugano, Switzerland, 2019, pp. 1–10.

[13] B. Lehner, K. Koutini, C. Schwarzlmüller, T. Gallien, and G. Widmer, "Acoustic scene classification with reject option based on ResNets," in *Proc. Detection Classification Acoustic Scenes Events Workshop (DCASE)*, New York, NY, USA, 2019, pp. 1–4.

[14] R. N. D'souza, P.-Y. Huang, and F.-C. Yeh, "Structural analysis and optimization of convolutional neural networks with a small sample size," *Sci. Rep.*, vol. 10, no. 1, pp. 1–13, Dec. 2020.

[15] E. Sprengel, M. Jaggi, Y. Kilcher, and T. Hofmann, "Audio based bird species identification using deep learning techniques," CLEF, Évora, Portugal, Tech. Rep. Vol-1609, 2016.

[16] Z. Zhu, H. Wang, T. Zhao, Y. Guo, Z. Xu, Z. Liu, S. Liu, X. Lan, X. Sun, and M. Feng, "Classification of cardiac abnormalities from ECG signals using SE-ResNet," in *Proc. Comput. Cardiology Conf. (CinC)*, Rimini, Italy, Dec. 2020, pp. 1–4.

[17] P. Stoica and R. L. Moses, *Spectral Analysis of Signals*. Upper Saddle River, NJ, USA: Prentice-Hall, 2005.

[18] M. Thomas, B. Martin, K. Kowarski, B. Gaudet, and S. Matwin, "Marine mammal species classification using convolutional neural networks and a novel acoustic representation," 2019, *arXiv:1907.13188*.

[19] K. M. Stafford, C. G. Fox, and B. R. Mate, "Acoustic detection and location of blue whales (Balaenoptera musculus) from SOSUS data by matched filtering," *J. Acoust. Soc. Amer.*, vol. 96, no. 5, pp. 3250–3251, Nov. 1994.

[20] J. Allen and H. Gordon. (2007). *Watkins Marine Mammal Sound Database*. Woods Hole Oceanographic Institution Marine Mammal Center. Accessed: May 23, 2021. [Online]. Available: https://cis.whoi.edu/science/B/whalesounds/index.cfm

[21] T.-H. Lin and L.-S. Chou, "Automatic classification of delphinids based on the representative frequencies of whistles," *J. Acoust. Soc. Amer.*, vol. 138, no. 2, pp. 1003–1011, Aug. 2015.

[22] S. L. Marple, *Digital Spectral Analysis With Applications*. Upper Saddle River, NJ, USA: Prentice-Hall, 1987.

[23] D. Lyon, "The discrete Fourier transform, Part 4: Spectral leakage," *J. Object Technol.*, vol. 8, pp. 23–24, Nov. 2009.

[24] Y. Qu, X. Li, Z. Qin, and Q. Lu, "Acoustic scene classification based on three-dimensional multi-channel feature-correlated deep learning networks," *Sci. Rep.*, vol. 12, no. 1, pp. 1–11, Aug. 2022.

[25] K. Xu, D. Feng, H. Mi, B. Zhu, D. Wang, L. Zhang, H. Cai, and S. Liu, "Mixup-based acous-Mixup-based acoustic scene classification using multi-channel convolutional neural network," in *Proc. Pacific Rim Conf. Multimedia*, 2018, pp. 14–23.

[26] A. Ramdoss and P. Chithra, "Role of windowing techniques in speech signal processing for enhanced signal cryptography," in *Advanced Engineering Research and Applications*. Delhi, India: Research India Publications, 2017, pp. 446–458.

[27] A. R. Jayan, P. S. R. Bhat, and P. C. Pandey, "Detection of burst onset landmarks in speech using rate of change of spectral moments," in *Proc. Nat. Conf. Commun. (NCC)*, Bangalore, India, Jan. 2011, pp. 1–5.

[28] P. P. Das, S. M. Allayear, R. Amin, and Z. Rahman, "Bangladeshi dialect recognition using mel frequency cepstral coefficient, delta, delta-delta and Gaussian mixture model," in *Proc. 8th Int. Conf. Adv. Comput. Intell. (ICACI)*, Chiang Mai, Thailand, Feb. 2016, pp. 359–364.

[29] J. H. L. Hansen and B. D. Womack, "Feature analysis and neural network-based classification of speech under stress," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 4, pp. 307–313, Jul. 1996.

[30] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Highway networks," 2015, *arXiv:1505.00387*.

[31] A. Saffari, M. Khishe, and S.-H. Zahiri, "Fuzzy-ChOA: An improved chimp optimization algorithm for marine mammal classification using artificial neural network," *Anal. Integr. Circuits Signal Process.*, vol. 111, no. 3, pp. 403–417, Jun. 2022.

[32] A. Theissler, M. Thomas, M. Burch, and F. Gerschner, "Confusion-Vis: Comparative evaluation and selection of multi-class classifiers based on confusion matrices," *Knowl.-Based Syst.*, vol. 247, Jul. 2022, Art. no. 108651.

[33] I. Ryazanov, A. T. Nylund, D. Basu, I.-M. Hassellöv, and A. Schliep, "Deep learning for deep waters: An expert-in-the-Loop machine learning framework for marine sciences," *J. Mar. Sci. Eng.*, vol. 9, no. 2, p. 169, Feb. 2021.
[34] L. Sayigh, M. A. Daher, J. Allen, and E. al, "The watkins marine mammal sound database: An online, freely accessible resource," *Proc. Meetings Acoust.*, vol. 27, no. 1, pp. 1–9, 2016.
[35] M. Sankupellay and D. Konovalov, "Bird call recognition using deep convolutional neural network, ResNet-50," in *Proc. Acoustics*, 2018, pp. 1–8.

DANIEL T. MURPHY received the M.S. degree in computer science from The University of New Orleans, New Orleans, LA, USA, in 2021. From 2017 to 2021, he worked for the Gulf States Center for Environmental Informatics (GulfSCEI). He currently works as a Software Engineer at Cubrc Inc., Buffalo, NY, USA. His current research interests include machine learning and bioacoustic analysis.

ELIAS IOUP received the Ph.D. degree in engineering and applied science from The University of New Orleans. He is currently a Computer Scientist and the Head of the Geospatial Computing Section, U.S. Naval Research Laboratory. His research interests include high-performance geospatial data processing, geospatial and environmental web services, and geospatial data visualization.

MD TAMJIDUL HOQUE received the Ph.D. degree in information technology from Monash University, Melbourne, VIC, Australia, in 2008.

From 2007 to 2011, he was a Research Fellow with Griffith University, Brisbane, QLD, Australia. From 2011 to 2012, he was a Postdoctoral Fellow with Indiana University–Purdue University Indianapolis, Indianapolis, IN, USA. He is currently an Associate Professor with the Computer Science Department, The University of New Orleans, New Orleans, LA, USA. His current research interests include deep/machine learning, evolutionary computation, and artificial intelligence, applying toward hard optimization problems, especially for bioinformatics problems such as protein structure-prediction, disorder predictor, and energy function.

MAHDI ABDELGUERFI is currently a Professor and the Chairperson of the Computer Science Department, The University of New Orleans. He is also the Founder and the Executive Director of the Canizaro Livingston Gulf States Center for Environmental Informatics (GulfSCEI).