JOURNAL OF
EMERGING INVESTIGATORS

Article

Optimizing data augmentation to improve machine learning accuracy on endemic frog calls

Nemai Anand[1], Anand Sampath[2]

[1] University High School, Irvine, California

[2] Endemic Greens, Mudigere, Karnataka, India

SUMMARY

The mountain chain of the Western Ghats on the Indian peninsula, a UNESCO World Heritage site, is home to about 200 frog species, 89 of which are endemic. Distinctive to each frog species, their vocalizations can be used for species recognition. Manually surveying frogs at night during the rain in elephant and big cat forests is difficult, so being able to autonomously record ambient soundscapes and identify species is essential. An effective machine learning (ML) species classifier requires substantial training data from this area. The goal of this study was to assess data augmentation techniques on a dataset of frog vocalizations from this region, which has a minimal number of audio recordings per species. Consequently, enhancing an ML model's performance with limited data is necessary. We analyzed the effects of four data augmentation techniques (Time Shifting, Noise Injection, Spectral Augmentation, and Test-Time Augmentation) individually and their combined effect on the frog vocalization data and the public environmental sounds dataset (ESC-50). The effect of combined data augmentation techniques improved the model's relative accuracy as the size of the dataset decreased. The combination of all four techniques improved the ML model's classification accuracy on the frog calls dataset by 94%. This study established a data augmentation approach to maximize the classification accuracy with sparse data of frog call recordings, thereby creating a possibility to build a real-world automated field frog species identifier system. Such a system can significantly help in the conservation of frog species in this vital biodiversity hotspot.

INTRODUCTION

According to the United Nations Educational, Scientific, and Cultural Organization (UNESCO), predating the Himalayan mountains, the Western Ghats Mountain chain extends for approximately 1600 km (990 miles), encompassing high montane forest ecosystems that play a pivotal role in shaping the Indian monsoon weather patterns. This region has exceedingly rich biodiversity and endemism. It is designated among the top eight 'hottest hotspots' of biodiversity, harboring a minimum of 325 globally endangered species (1). Concerningly, when examining the growth of human population trends within biodiversity hotspots, the Western Ghats is one of the top three hotspots with the most substantial risk of biodiversity loss, based on human population density (2). Human activity within the Western Ghats spans over 12,000 years, with the last two centuries witnessing extensive logging and clearing for plantation crops (3).

Amphibians, including frogs, exhibit a heightened sensitivity to environmental shifts owing to their unique life cycle, characterized by an aquatic larval phase that transitions into a terrestrial adult stage (4). Surveying frogs in the Western Ghats poses inherent challenges due to their vocalization activity patterns occurring from early to late evening. The frogs advertise their presence in the dense canopy by vocalizing for several hours into the night. These calls help them to establish territories and to attract mates (5). Surveying frogs in a habitat consisting of dense foliage, steep slopes, high leech activity, venomous snakes, and large mammals require highly trained personnel and is difficult to scale.

Notably, frog advertisement calls play a crucial role as a pre-mating isolation mechanism, showcasing a high degree of species specificity (6). Automated identification of frog species based on their vocalizations can significantly increase the survey coverage of the habitat. Researchers have explored a variety of techniques to perform automatic animal species identification, including breaking the vocalization into a set of organized sequences of brief sounds from a species-specific vocabulary and categorizing calls into recognizable syllables (7). Others have looked at improving automatic recognition using temporal context inherent in vocalizations by means of a convolutional neural network (CNN) (8). It has also been shown that the inclusion of temporal information like the number of repetitions of certain call characteristics improves the automatic recognition and transcription of wildlife recordings (8).

Training a machine learning (ML) model for automatic recognition of frog vocalizations requires large training datasets, needing highly intensive efforts over extended durations of time in the monsoon season when most vocalizations take place. Training datasets of twenty thousand samples per class have been used to recognize bird and whale sounds (8). However, in the case of rare species, due to their scarce prevalence, obtaining multiple recordings from different individuals from multiple locations is difficult, but possible. For frog vocalizations in the Western Ghats region of India, publicly available data sources at most contain a few species with single individual recordings (9). For smaller datasets, data augmentation can generate new samples by perturbing the data and then adding the new samples to the original data to expand the dataset. This is considered a regularization method by increasing the diversity of training data (10). Given the difficulty in collecting training data with various endemic frog species from this environment, maximizing the classification accuracy with limited training data sets using data augmentation becomes a prerogative. Therefore, we examined the impact of suitable data augmentation methods on frog vocalization data to determine model accuracy.

We hypothesized that data augmentation would improve the ML model's accuracy on the frog vocalization dataset. To test our hypothesis, frog vocalizations at the site of Endemic Greens in the Mudigere region of Western Ghats were collected. We evaluated the effects of augmentation methods on the combined training and test methods by analyzing the Convolutional Neural Network's performance metrics. We found that the efficacy of augmentation techniques varies based on the data collection process and that different techniques are affected to different extents for the same change in the data collection process. This finding allows the determination of the data collection parameters including the number of recordings per frog species, noise effects, recording length needed to effectively perform an automatic species identification in the field, and the corresponding augmentation techniques that work best with the collected data. With these parameters the significant human effort to collect the field recordings under challenging conditions can be optimized to allow more endangered frog species to be surveyed, which in turn will allow for a better biomarker of the biodiversity health in the ecologically sensitive Western Ghats region.

RESULTS

To identify the frog species based on the recorded calls from the Western Ghats, an expert naturalist and zoologist trained on the calls from endemic and local frog species identified the species based on the recorded frog calls from this location (**Figure 1**).

We applied three data augmentation methods on the training data: Spectral, Time Shift and Noise Injection (11). We also ran the neural network on data without any augmentation as a control (**Figure 2A**). Spectral Augmentation works in the frequency and time domain and randomly obstructs the spectrogram by masking frequency values (**Figure 2B**). Time Shifting creates augmented samples in the time domain and randomly changes the tempo and length of the audio sample without changing the frequency component (**Figure 2C**). Noise injection adds Gaussian noise at random noise levels to the original sample to create the augmented samples (**Figure 2D**). The effects of each training data set augmentation method were evaluated both independently and when all methods were combined.

Data augmentation methods on the test data set utilized a single approach, the Test-Time augmentation method. Test-

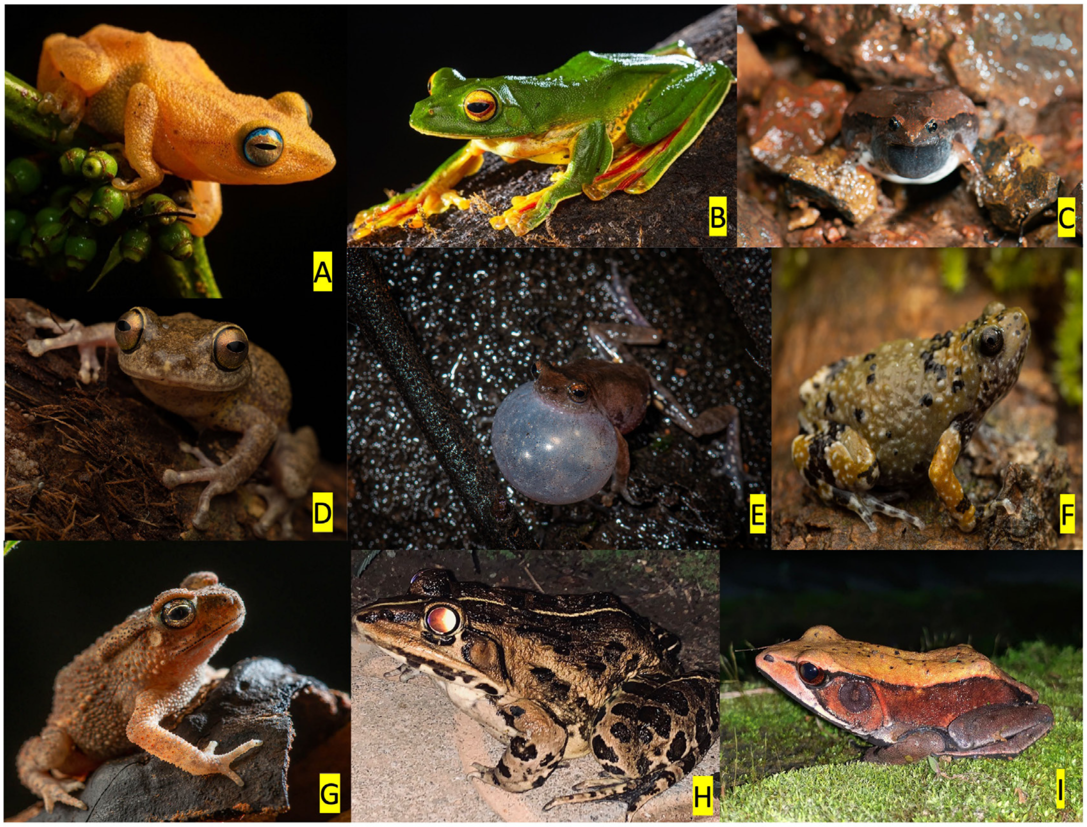

Figure 1: Western Ghats frogs The frogs of the Western Ghats whose advertisement calls were recorded and used in the FROGS-WGHATS data set. A total of 12 frog species had multiple calls recorded and of the 12 species, only 9 have been photographed. (A) Blue Eyed Bush Frog - *Raorchestes luteolus*, (B) Malabar Gliding Frog - *Rhacophorus malabaricus*, (C) Ornate Narrow Mouthed Frog – *Microhyla ornata* (D) Rao's Intermediate Golden Backed Frog- *Hylarana intermedius* (E) Bombay Bush Frog - *Raorchestes bombayensis* (F) Indian Dot Frog – *Ramanella mormorata* (G) Ferguson's Toad – *Duttaphrynus scaber* (H) Bull Frog- *Hoplobatrachus tigrinus* (I) Bi-colored Frog – *Clinotarsus curtipes*.

https://doi.org/10.59720/23-223

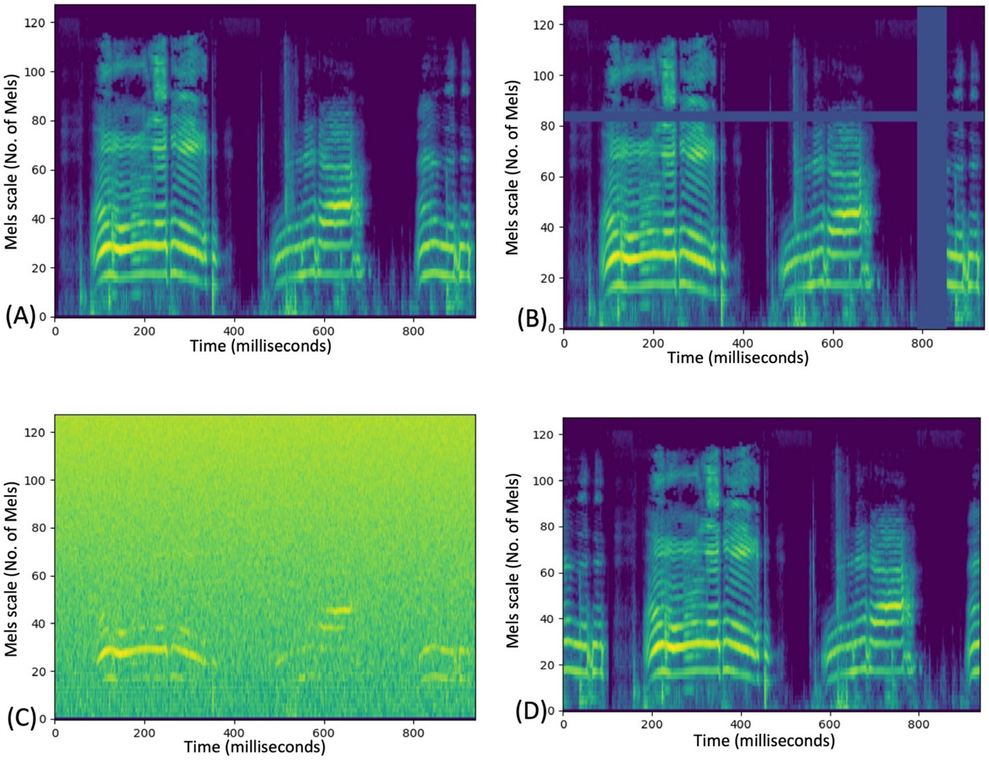

Figure 2: Sample audio mel spectrogram from the ESC-50 Dataset. Raw audio data was transformed into a Mel spectrogram with the following hyperparameters and augmentation techniques applied: 64 Mels, a 256 hop length, and the number of FFTs was set to 1024 (A) No Augmentation (B) Spectral Augmentation (C) Noise Injection (D) Time Shift. Augmentation techniques C and D were performed on the audio data before it was transformed into its spectrogram form using torchaudio.transforms.MelSpectrogram() while technique B was done after.

Time augmentation expands the testing data set and employs the Spectral augmentation method on this data set (12). Since Test-Time augmentation is applied only to the testing data set, the effect of this method is evaluated independently of the methods used on the training data sets.

These augmentation techniques were applied to three different datasets. One is a dataset that comprises of very small amount of audio samples from just endemic frogs from the Western Ghats region, this will be referred to as FROG_WGHATS. The other two datasets are derived of an online environmental sound dataset called ESC-50 which contains 50 different categories of environmental sounds. The ESC-50-SMALL contains portions of the audio samples from the entire dataset while ESC-50-FULL contains all the samples in the original dataset (13). All these three datasets have different sizes, and that may be a factor that affects the performance of the model.

We determined the accuracy of each of the data augmentation methods and compared it to the baseline accuracy that was obtained without any data augmentation. This measure was then used to obtain the relative accuracy improvement for a single or combined group of data augmentation techniques that were applied to the dataset. Due to the potential variability of augmentation methods, a set of 10 trials were run on each data augmentation method for a given dataset. The average accuracy across the 10 trials was then used as the measure of the classification accuracy for the augmentation(s) technique along with the relative accuracy improvement compared to the baseline (**Table 1**).

We also looked at the "F-score", which combines the precision and recall by determining their harmonic mean. Precision and recall are useful when the impact of a false positive or false negative is more significant on the model's objective. For the environmental sound classification, the weight is similar meaning that F-score (the combination of both) is a good metric to use when testing the accuracy. The F-score was micro-averaged across the classes and used to confirm the trend in accuracy.

F-scores and accuracy had similar trends with an R^2 value of 0.8 (**Figure 3**). Consequently, the F-score test allows the use of the accuracy and the relative accuracy metric to confirm the hypothesis in the study. On the ESC-50-FULL dataset, the baseline, un-augmented accuracy was 31%. The relative accuracy improvement for Time Shift, Noise Injection, Spectral Augmentation, and Test-Time augmentation was 29%, 28%, 30%, 23% respectively (**Figure 4**). With all four augmentation methods on the ESC-50-FULL dataset, the relative accuracy improvement was 38%, and without Test-Time augmentation, the relative accuracy improvement was 34% (**Figure 4**).

On the ESC-50-SMALL dataset, the baseline, un-augmented accuracy was 22%, and the relative accuracy improvement for Time Shift, Noise Injection, Spectral Augmentation, and Test-Time augmentation was 21%, 20% 23%, and 29%, respectively. With all four augmentation methods on the ESC-50-SMALL dataset, the relative accuracy improvement was 49%, and without Test-Time augmentation, it was 44% (**Figure 4**).

On the FROG-WGHATS dataset, the baseline, un-augmented accuracy was 8%, and the relative accuracy

https://doi.org/10.59720/23-223

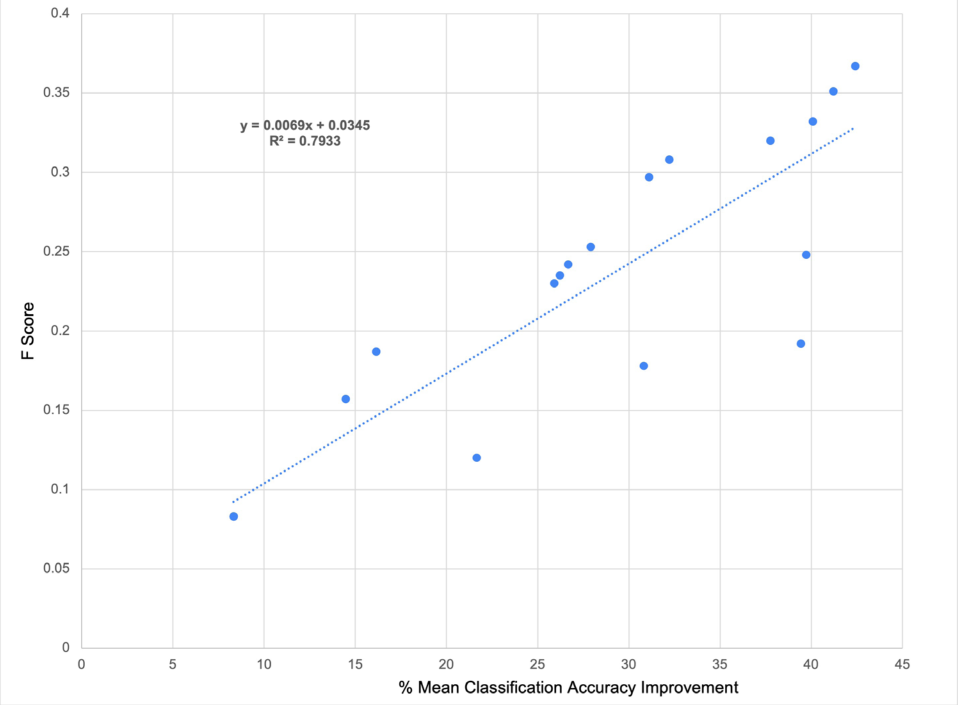

Figure 3: F-score vs classification accuracy. This figure displays the linear relationship between the average classification accuracy improve of 10 trials for each combination of augmentation techniques applied on the CNN and the corresponding F-scores which were also averaged. The accuracy was derived from the predictions while the F-score was calculated from using sklearn.metrics.f1_score(). The coefficient of determination is .7933 indicating that 79.33% of the variation accounted for in the F Score can be explained by the Mean Classification Accuracy Improvement

improvement for Time Shift, Noise Injection Spectral augmentation, and Test-Time augmentation was 0%, 0%, 0%, and 74%, respectively. With all four augmentation methods on the FROG-WGHATS dataset, the relative accuracy improvement was 94%, and without Test-Time augmentation, it was 0%. A limitation is that while the relative accuracy significantly improved, the overall accuracy was still fairly low at 16% (**Figure 4**).

DISCUSSION

We sought to evaluate the effects of the CNN's classification accuracy with data augmentation on the frog vocalizations dataset, and we hypothesized that data augmentation would bolster the performance of the CNN. The augmentation techniques were concurrently applied to a known reference sound data set to verify the proper application of the augmentation methods, serving as an experimental 'control'. The reference dataset has 40 recordings per sound origin. Additionally, the same reference dataset was then trimmed to have a 'SMALL' version which contains 15 recordings per sound origin. Each unique sound origin within the ESC-50 such as a Dog, a Crow or an Insect is known as a class within the data set and each class has multiple recording samples. In the FROG-WGHATS dataset each class represents a unique frog species with multiple recordings. We analyzed the three datasets, which had different amounts of environmental sound recordings in their test sets. ESC-50-FULL with 40 recordings per class, ESC-50-SMALL with 15 recordings per class, and FROG-WGHATS with 1-3 recordings per class. Data augmentation methods were applied to the training and

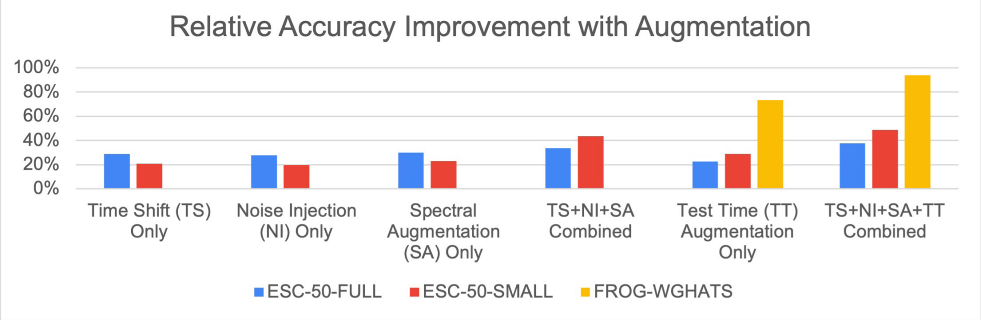

Figure 4: Relative Accuracy Improvement of Data Augmentation. This graph shows all sets of augmentation techniques that were tested on the CNN and their accuracies across the 3 datasets. CNN run 6(10) times where 6 represents the number of subsets of augmentation tests and 10 represents the number of trials. For every subset of augmentation tests, the data is either augmented or inputted as normal depending on what is being tested.

https://doi.org/10.59720/23-223

test data in accordance with the combination of augmentation techniques were being tested. The combined augmentation effect on both the training and test data was also evaluated.

We examined the effectiveness and efficiency of the 3 augmentation techniques done on the training set. On the ESC-50 FULL and ESC-50 SMALL datasets, Spectral Augmenting produced a relative increase in accuracy by 30% and 23%, respectively. Time Shifting increased the accuracy by 29% and 21%, respectively. Noise Injection increased the accuracy by 28% and 20%, respectively. The Spectral Augmenting randomly masked values on both dimensions (frequency and time). Performing this operation directly on the spectrogram created data variance through significant visual obstruction because it dramatically altered the appearance of a feature, given that the random masking successfully covered features in a transformation. A possible reason why Spectral Augmentation was the most successful technique could be due to the dataset not already having any masking of values, so the variance added by spectral augmentation is unique. When Time Shifted, the audio data (which contains amplitude data sampled at a standardized rate) will be shifted by a specified time instance. This efficiently covers up and removes features in the audio data again through a visual obstruction, and it eliminates prominent or insignificant features, which increases the weights of other features. Time Shifting was shown to be very effective, but a dataset can account for this variance naturally because it's unnatural to have multiple recordings that are almost the exact same. Furthermore, the Noise injection method was based on a random noise level, which was synthetically generated and combined with the raw input sound. The noise level is a randomly chosen value between the range of 0 and 1, which acts as a scalar for the noise. The noise is an array of samples from a standard normal distribution (mean 0, variance 1). The product of the noise level and the noise was added to the raw input sound for modification. Noise injection may not have produced as much of a significant relative accuracy increase because of the data collection process. Data was collected during times when there were many conflicting sounds and not only the designated frog sound, meaning the dataset most likely accounts for background noise already. This may be the cause of Noise Injection's lower performance.

Using all four techniques led to a remarkable enhancement in the ML model's classification accuracy for the Frog calls dataset by 94% and by 38% and 49% for the ESC-50-FULL and ESC-50-SMALL datasets respectively. Furthermore, augmenting test data creates a better representative test set and performs better on the smallest testing sizes. On the three datasets: ESC-50-FULL, ESC-50-SMALL, and FROG-WGHATS, test-time augmentation alone improved the accuracy significantly. It made the model's performance go up by 23%, 29%, and 74% for each dataset. When we combined this with the other three augmentation techniques, the effects were even larger – 38%, 49%, and 94%, respectively. Test-time augmentation works by subjecting the test data to the spectral augmentation method synthetically, increasing the number of samples and variance in the testing data set. The effectiveness of the Test-time augmentation increases as the number of available samples in the test set decreases. This inverse relationship can be explained by the lack of sufficient variance in the test data set and the Test-time augmentation increases this variance.

A pivotal insight from this study is that with the appropriate choice of data augmentation techniques, a small number of samples can be used to improve the classification accuracy of frog vocalizations. Since data augmentation introduces unique perturbations and variance to the training datasets, the combined effect of augmentation methods on training and test datasets performs better on smaller-sized datasets. Another noteworthy takeaway is the number of samples needed in a testing set to gauge the model's performance accurately. After noticing a significant improvement in a model's performance after implementing Test-Time data augmentation, it's clear the original testing set was too small to fairly determine the model's performance allowing us to determine a minimum sample size. Notably, the study identifies that a robust field-applicable classifier can be crafted with about 15 data samples per class. This finding serves as a crucial threshold for guiding future data collection endeavors within the Western Ghats to build a field-deployable autonomous frog species identifier.

Based on these findings, the most suitable augmentation technique will depend on its impact on the data and the hyperparameters used. For datasets that have low noise, it may be better to use noise injection more than other methods, whereas for datasets with samples that have similar timestamps, Time Shifting may be more effective. Spectral augmenting can then be stated as an augmentation technique that is independent of a dataset, since it can be very difficult to alter its performance because it is difficult to naturally mask values. Methods like Time-Shifting and Noise Injection can have a reduced effect because of the already present amounts of noise in the dataset and the differences in times voice recordings are taken, showing that their efficacy can be dependent. However, it's difficult to mask frequency values in audio samples naturally, so Spectral Augmenting's efficacy remains more static throughout a variety of datasets, making its performance independent. For this reason, spectral augmenting is most likely the most effective and efficient technique to use for audio processing. Furthermore, when different augmentation techniques are used together their effects add up because of their unique perturbations. This is generally untrue when it comes to increasing the number of samples one specific augmentation technique is augmenting.

A source of error is mostly likely the type of classification that is being performed on these audio files. In the ESC-50 FULL and ESC-50 SMALL datasets, the classification problem involves discerning different environmental sounds (Dogs, roosters, pigs, rain, fire crackling, fireworks, helicopters, trains, etc.). The FROG-WGHATS dataset comprises frog vocalizations that are in a specific ecosystem. Classifying different environment sounds may be easier than classifying different frog species due to the variation in the feature extraction process, meaning that there are more prominent distinctions in different organisms' sounds compared to distinctions within a population of a given species. Another limitation is the difference in the number of classes. Two datasets have 50 classes, while the other dataset has 12 classes. Classification done on fewer classes is generally easier because the random probability of guessing a class is higher when the number of classes is lower.

There are many more augmentation techniques that are done for signal processing on environmental sounds than the ones that were explored in the study. For example, one can perform Pitch Shift (frequency is randomly modified), as

https://doi.org/10.59720/23-223

well as a Time stretch (sound is randomly slowed or sped up) (14). Analyzing the results of the combination of multiple augmentation techniques of a deep convolutional neural network provides interesting insights. One can also analyze the extent to which the combination of augmentation techniques will increase a model's performance, but it will be important to examine the dataset to identify if certain techniques are dependent on the audio data and identify the techniques that will provide the best results. The augmentation techniques that are independent of the data should be used.

MATERIALS AND METHODS

The comprehensive ESC-50 dataset (ESC-50-FULL) encompasses 50 distinct classes, each containing 40 recordings of environmental sounds. The online public datasets called ESC-50 were downloaded from the public Github repository called ESC-50, created by Karol Piczak(13). Additionally, a reduced subset of ESC-50 (ESC-50-SMALL) was considered, maintaining the same 50 classes but comprising only 15 recordings per class.

The frog vocalizations dataset from the Western Ghats (FROG-WGHATS) consisted of 12 classes with 1-3 recordings per class. Each of the 12 classes represented a unique frog species. 28 individual recordings were collected using a commercially available field audio recorder F3 (Zoom) audio recorder with an MKE 600 (Sennheiser) shotgun microphone with a sampling frequency of 44K Hz. The recordings were made in the central region of the Western Ghats between January and June of 2023. The species were identified from the calls by an expert naturalist who is trained on recognizing frog calls, confirming the species visually and, when possible, with a photographic record. Pictures of a subset of the species were also obtained during this period. The frogs' vocalizations were naturally occurring in their native habitat and frogs were not disturbed.

The following steps apply to all three datasets. After the data was collected, Python-based audio processing and machine learning libraries were downloaded and imported for use: Pytorch, Numpy, Matplotlib, and Pandas (15-18). Once the paths were loaded using pathLib, the raw audio data was loaded using Pytorch. The data was split into a training-testing ratio of 8:2. The audio data was pre-processed to meet the requirements to be trained and tested on a model. The data was rechanneled from whichever channel it was into stereo. Then it was sampled to a standardized sampling rate. When audio is initially collected, it is possible that recordings were sampled at different rates, and this can lead to files having more or less data although the recordings are of the same length. Standardizing the sampling rate will make sure that all recordings of the same time frame have the same length so they can be processed by the model. The data was padded with 0s or truncated in length to ensure that there is constant dimensionality in the data, and it can be processed by the model. Using previously established functions found online, the noise injection data augmentation function and the time shift data augmentation function were defined. These functions were used to augment the raw audio data. The parameters of the augmentation methods that affect these data were randomized. The hyperparameters used for Time-Shift were a shift limit of 6 times the length of the audio sample. The process involved randomly circulating the data by a value between (0 and 6 times the length of the audio sample). When Time-Shifted, the audio data (which contains amplitude data sampled at a standardized rate) will be shifted by a specified time instance. The hyperparameters for noise injection consisted of the noise level. The Noise level had random values from 0 to 1, and the generated noise was randomly generated with normally distributed values with a mean of 0 and a standard deviation of 1. The raw input sound contains values ranging from -1 to 1, so the impact of the Noise injection would be significant. The raw audio data was converted into a Mel spectrogram with the following hyperparameters: 64 Mels, a 256 hop length, and the number of FFTs was set to 1024. Then the spectral augmenting function was created and defined based on a previously established method found online. Masking was performed on both the time and frequency axis with 1 masking per axis. The width of each masking was randomized. The test-time augmentation function was defined, which employed spectral augmentation on the testing datasets. After all the augmentations are performed, the labels were updated to match the length of the data.

The CNN model (19) used had a total of 11 layers, the first layer was a Conv2D layer with 32 kernels and has a filter size of 3x3, taking in an input shape of 64, 938, 2. Rectified Linear Unit (Relu) was used as the activation function. Next, there is a max pooling layer with a 2x2 filter size. The next 6 layers contain stacked Conv2D and 3 max-pooling layers. The filter sizes for all the Conv2D layers were 3x3 and all the filter sizes for the max-pooling layers was 2x2. The first Conv2D layer has 64 kernels, the second has 64 kernels, and the third has 128 kernels. All the Conv2D layers have Relu as their activation function. The next two layers consist of a Flattening layer and a Dense layer with 64 neurons and a Relu activation function. The final layer is a Dense layer with 50 neurons and softmax activation. When testing the model on the FROG-WGHATS dataset it's basically the same except the final layer has 12 neurons instead of 50. The preprocessing techniques can be applied to all the datasets, it is just the final layer of the model that differs when testing. The model used the Adam optimizer, the Sparse Categorical CrossEntropy loss function, and accuracy as a metric. For each set of augmentation techniques, the model was run until the rate of change of the testing loss was not decreasing. The effectiveness of each augmentation technique was determined by comparing the results on each of the datasets to identify how differences in datasets affect the relative performance of the different technique.

ACKNOWLEDGMENTS

We would like to acknowledge Raju A.K, Operations Director and Naturalist of Endemic Greens, and his local team for collecting the frog call recordings and images along with species classifications, and they provided us the rights to use these materials exclusively for this study. We would like to thank Prashanth Chandran, Vice President, Artificial Intelligence, Masimo for the machine learning guidance.

Received: August 22, 2023
Accepted: March 8, 2024
Published: March 09, 2025

REFERENCES

1. "Western Ghats." *UNESCO World Heritage Convention*,

whc.unesco.org/en/list/1342/. Accessed 6 Jul 2024.

2. Cincotta, Richard et al. “Human population in the biodiversity hotspots.” *Nature,* Vol 404, 2000, pp 990–992, https://doi.org/10.1038/35010105.
3. M.O. Anand et al. “Sustaining biodiversity conservation in human-modified landscapes in the Western Ghats: Remnant forests matter.” *Biological Conservation*, Volume 143, Issue 10, 2010, pp. 2363-2374, ISSN 0006-3207, https://doi.org/10.1016/j.biocon.2010.01.013.
4. Rowley, JJL et al. “The FrogID dataset: expert-validated occurrence records of Australia’s frogs collected by citizen scientists.” *Zookeys,* vol. 912, pp. 139-151, https://doi.org/10.3897/zookeys.912.38253.
5. Shikhara, Ananda. “Croaking to the Audience.” From the Field, jlrexplore, June 01, 2022, jlrexplore.com/explore/from-the-field/croaking-to-the-audience.
6. Blair, WF, “Isolating Mechanisms and Interspecies Interactions in Anuran Amphibians” *The Quarterly Review of Biology*, vol. 39, pp. 334-44, Dec 1964. https://doi.org/10.1086/404324.
7. Chenn-Jung Huang et al. “Applications of data mining techniques to automatic frog identification”, *Applied Artificial Intelligence*, vol. 23, pp. 553-569, July 2009. https://doi.org/10.1080/08839510903145223
8. Madhusudhana S et al. “Improve automatic detection of animal call sequences with temporal context.” *J R Soc Interface*. vol. 18(180), July 2021, https://doi.org/10.1098/rsif.2021.0297.
9. Ramya. B et al, “Mandookavani – An Acoustic Guide to the Frogs and Toads of the Western Ghats.”, Conservation India, May 01, 2015, www.conservationindia.org/resources/mandookavani-an-acoustic-guide-to-the-frogs-and-toads-of-the-western-ghats
10. Shengyun, Wei et al, “A Comparison on Data Augmentation Methods Based on Deep Learning for Audio Classification.” *Journal of Physics:Conference Series*, vol. 1453, October 2019, https://doi.org/10.1088/1742-6596/1453/1/012085.
11. Herrera, Jose, “Audio Data Augmentation: Techniques and Methods.” pangeanic, Pangeanic, 06/16/23, blog.pangeanic.com/audio-data-augmentation-techniques-and-methods
12. Kimura, Masanari, “Understanding Test-Time Augmentation.” arxiv, Arxiv, 02/10/24, arxiv.org/html/2402.06892v1
13. Karol J.” ESC Dataset for Environmental Sound Classification”. *Proceedings of the 23rd ACM conference on Multimedia*, pp. 1015-1018. https://doi.org/10.1145/2733373.2806390. Github, github.com/karolpiczak/ESC-50.
14. Ma, Edward “Data Augmentation for Audio.” *Medium.Com*, 12 Jun. 2019 medium.com/@makcedward/data-augmentation-for-audio-76912b01fdf6
15. Paszke, Adam, et al. “PyTorch: An Imperative Style, High-Performance Deep Learning Library.” *33rd Conference on Neural Information Processing Systems*, edited by H. Wallach et al., Curran Associates, Inc. Red Hook NY, 2019, pp. 8024-8035.
16. Harris, Charles R., et al. “Array Programming with NumPy.” *Nature*, vol. 585, 2020, pp. 357-362. https://doi.org/10.1038/s41586-020-2649-2.
17. Hunter, John D. “Matplotlib: A 2D Graphics Environment.” *Computing in Science & Engineering*, vol. 9, no. 3, 2007, pp. 90-95. https://doi.org/10.1109/MCSE.2007.55.
18. McKinney, Wes. “Data Structures for Statistical Computing in Python.” *Proceedings of the 9th Python in Science Conference*, vol. 445, 2010, pp. 51-56 https://doi.org/10.25080/Majora-92bf1922-00a.
19. Anand, Nemai. “Github repository for python code used in this paper” github.com/blueflare743/FrogProject-V1.00-