

**SRI LANKAN ELEPHANT SOUND
CLASSIFICATION USING DEEP LEARNING**

Ariyasingha Gamage Hiruni Udarika Dewmini

219327R

Master of Science in Computer Science Specializing in Data
Science, Analytics and Engineering

Department of Computer Science & Engineering
Faculty of Engineering

University of Moratuwa
Sri Lanka

June 2024

SRI LANKAN ELEPHANT SOUND CLASSIFICATION USING DEEP LEARNING

Ariyasingha Gamage Hiruni Udarika Dewmini

219327R

Thesis submitted in partial fulfillment of the requirements for the degree
Master of Science in Computer Science Specializing in Data Science,
Analytics and Engineering

Department of Computer Science & Engineering
Faculty of Engineering

University of Moratuwa
Sri Lanka

June 2024

DECLARATION

I declare that this is my own work and this Thesis does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or Institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature:

Date:

The supervisor should certify the Thesis with the following declaration.

The above candidate has carried out research for the Master of Science in Computer Science Specializing in Data Science, Analytics and Engineering Thesis under my supervision. I confirm that the declaration made above by the student is true and correct.

Name of Supervisor: Prof. Dulani Meedeniya

Signature of the Supervisor:

Date:

ACKNOWLEDGEMENT

I extend my sincere gratitude to Professor Dulani Meedeniya for her invaluable guidance and unwavering motivation throughout this research endeavor. Additionally, I express profound appreciation to my parents, friends, and the esteemed members of the Department of Computer Science and Engineering for their steadfast support and invaluable assistance, which have been instrumental in the realization of this work.

ABSTRACT

Understanding elephant caller types is crucial for various aspects of wildlife conservation and ecological research. By decoding the intricate vocalizations of elephants, researchers gain valuable insights into their behavior, social dynamics, and emotional expressions, which are pivotal for species conservation efforts. Elephant vocalizations serve as indicators of ecosystem health and vitality, aiding in ecological monitoring and biodiversity conservation initiatives. Furthermore, investigating caller types contributes to the preservation of cultural heritage by honoring the profound connection between humans and elephants across generations. In essence, delving into the world of elephant communication not only advances scientific knowledge but also fosters harmony between humans and these majestic animals, ensuring their long-term survival in the wild.

In this study, we delve into the domain of elephant caller-type classification utilizing raw audio format processing. Our focus lies on exploring lightweight models suitable for deployment on edge devices, including MobileNet, YAMNET, and RawNet, alongside introducing a novel model termed ElephantCallerNet, based on ACDnet architecture. Notably, our investigation reveals that the ACDnet-based ElephantCallerNet achieves an impressive accuracy of 89% when applied to a raw audio dataset. Leveraging Bayesian optimization techniques, we fine-tune crucial parameters such as learning rate, dropout, and kernel size, thereby enhancing model performance. Moreover, we scrutinize the efficacy of spectrogram-based training, a prevalent approach in animal sound classification. Through comparative analysis, we ascertain that for our dataset, raw audio processing outperforms spectrogram-based methods. In contrast to other models in the literature that primarily focus on a single caller type or binary classification (such as identifying whether a sound is an elephant voice or not), our models are designed to classify three distinct caller types: Roar, Rumble, and Trumpet. This approach significantly increases the complexity of our experiments compared to those discussed in the literature.

In the domain of elephant vocalization analysis, there has been limited exploration into the direct processing of raw audio data. Predominantly, various feature extraction techniques have been employed before training machine learning algorithms. In our investigation, we aim to bypass preprocessing stages and directly input raw audio data into machine learning models to assess the feasibility and efficacy of training on unprocessed audio signals.

Keywords: Elephant Vocalizations, Supervised-learning, Raw Audio Processing, Feature extraction, Classification

TABLE OF CONTENTS

Declaration of the Candidate & Supervisor	i
Acknowledgement	iii
Abstract	v
Table of Contents	vii
List of Figures	xi
List of Tables	xv
List of Abbreviations	xvi
List of Appendices	xix
1 INTRODUCTION	1
1.1 Elephant Statistics in Sri Lanka	2
1.2 Elephant Vocalization Ranges	3
1.3 Problem Statement	4
1.4 Research Objectives	6
1.5 Research Questions	6
1.6 Research Scope	7
1.7 Importance of Processing Raw Audio	8
1.8 Limitations	8
2 LITERATURE REVIEW	11
2.1 Overview of Elephant Sounds Classification	11
2.2 Types of Elephant Vocalizations	12
2.3 Datasets	13
2.3.1 Asian Elephant Vocalizations Dataset	13
2.3.2 Elephant Listening Project (ELP)	14
2.3.3 ElephantVoices Dataset (www.ElephantVoices.org) [1]	14
2.4 Augmentation of Sound Data	14
2.4.1 Time Stretching	16
2.4.2 Pitch Shifting	17

2.4.3	Random Noise	17
2.4.4	Time Masking	19
2.4.5	Frequency Masking	19
2.5	Elephant Sounds Classification Using Feature Extraction Techniques	21
2.5.1	Feature Extraction Techniques	21
2.5.2	Related Studies with Feature Extraction Techniques	28
2.6	Elephant Sounds Classification Using Raw Audio Representation	31
2.6.1	Deep Learning Models for Raw Audio Classification	33
2.7	Taxonomy of Bioacoustic Audio Analysis Methods	40
2.8	Studies With Audio Classification	41
2.8.1	Classifying Environmental Sounds on Limited Devices: A Framework to Enhance the Performance of Deep Acoustic Networks in resource-limited environments [2]	41
2.8.2	Detecting Wild Elephants in Asia Through Advanced Frequency Domain Acoustic Analysis [3]	43
2.8.3	Animal Hunt: An AI-Powered Application for Recognizing Animal Sounds [4]	45
2.8.4	Creating a Convolutional Neural Network Stack, Ensemble for Recognizing Sounds from the Environment [5]	48
2.8.5	Exploring Elephant Sound Classification Through Parallel Convolutional Neural Networks [6]	48
2.8.6	Developing a Framework for Analyzing Bioacoustic Vocalizations With Hidden Markov Models[7]	50
2.9	Performance Matrices	54
2.9.1	Accuracy	54
2.9.2	Recall or Sensitivity	55
2.9.3	Precision	55
2.9.4	F1-score	55
2.9.5	Multi-class Confusion Matrix	55
3	METHODOLOGY	57
3.1	Overview of Audio Data Processing	57

3.2	Datasets	61
3.2.1	Identified Caller Types in Proposed Method	61
3.2.2	Data Preprocessing	62
3.3	Feature Extraction Techniques for Audio-Visual Representation	71
3.3.1	Mel-Frequency Cepstral Coefficients (MFCCs)	72
3.3.2	Chroma Feature	75
3.4	Implemented Machine Learning Algorithms	77
3.4.1	MobileNetV2	78
3.4.2	YAMNet	81
3.4.3	RawNet	84
3.4.4	ElephantCallerNet: ACDNet-Based Modified Raw Audio Model	86
3.4.5	ResNet18	93
3.5	Weight Initialization	95
3.5.1	Why He Initialization is Effective for Models Like RawNet and YAMNet	96
3.6	Loss Function	97
3.6.1	Cross-Entropy Loss	97
3.7	Optimization and Hyper-parameter Tuning With Bayesian Algorithm	98
3.8	Elephant Monitoring System Web Application	98
3.9	Implementation Aspects	101
4	RESULTS	103
4.1	Model Performance Comparison With Varying Dataset Sizes and Augmentation Levels: Smaller Dataset vs Larger Dataset	103
4.1.1	Performance Comparison Between Larger and Smaller Datasets for Raw Audio	103
4.1.2	Performance Comparison Between Larger and Smaller Datasets for Audio Visual Representation	105
4.2	Results Analysis of Raw Audio Classification Using Smaller Dataset	109
4.2.1	Model Size Comparison Among Implemented Models	110
4.2.2	Performance Metrics for Each Caller Type Across Implemented Models	113

4.2.3	Inference Time for Implemented Models	116
4.3	Results of Individual Models on Raw Audio	116
4.3.1	Results for MobileNetv2 Model	117
4.3.2	Results for YAMNet Model	120
4.3.3	Results for RawNet Model	121
4.3.4	Results for ElephantCallerNet Model	124
4.4	Results Analysis of Spectrogram Based Classification: MFCC, Chroma CQT feature extraction	126
4.5	Raw Audio Processing Vs Audio Spectrogram Processing	131
5	DISCUSSION	137
5.1	Study Contribution	137
5.1.1	Novel Contributions - ElephantCallerNet	138
5.1.2	ElephantCallerNet for Smaller Dataset	140
5.1.3	Generalizability of ElephantCallerNet	141
5.2	Comparison With Traditional Spectrogram Based Methods	147
5.3	Application of Research Findings	147
5.4	Future Directions	149
5.5	Experimental Limitations and Challenges	150
6	CONCLUSION	155
	References	157
	Appendix A Accoustic Features for Each Elephant Caller Type	165
	Appendix B Code Bases	171

LIST OF FIGURES

Figure	Description	Page
Figure 1.1	The Current Elephant Distribution. Retrieved from [8]	2
Figure 2.1	Data Augmentation Approaches from Surveyed Articles from [9]	16
Figure 2.2	Summary of Data Augmentation Methods on Animal Vocalizations. Retrieved from [9]	16
Figure 2.3	Comparison Between Data Augmentation Techniques. Retrieved from [10]	20
Figure 2.4	Feature extraction techniques used in bioacoustics. Retrieved from [11]	22
Figure 2.5	MFCC Coefficient Workflow	23
Figure 2.6	LPC Coefficient Workflow	24
Figure 2.7	Spectral sub-band centroids calculation steps	24
Figure 2.8	MFCC, Chroma, RMS, and Spectral Contrast Features for Roar	26
Figure 2.9	MFCC, Chroma, RMS, and Spectral Contrast Features for Rumble	27
Figure 2.10	MFCC, Chroma, RMS, and Spectral Contrast Features for Trumpet	27
Figure 2.11	Classification algorithms used for bioacoustics. Retrieved from [11]	29
Figure 2.12	SincNet Model Summary. Retrieved from [12]	34
Figure 2.13	Visualization of WaveNet Model with Dilated Causal Convolutional Layers. Retrieved from [13]	35
Figure 2.14	Visualization of WaveNet Model with Residual Block. Retrieved from [13]	36
Figure 2.15	YAMNet Model Architecture. Adapted from [14]	37
Figure 2.16	RawNet Model Architecture. Retrieved from [15]	38
Figure 2.17	MobileNet v2 Residual Connection. Retrieved from [16]	39
Figure 2.18	MobileNet v3 Connection. Retrieved from [17]	39
Figure 2.19	Taxonomy of Feature Extraction Techniques	41
Figure 2.20	Taxonomy of Model Training Techniques	42
Figure 2.21	ACDNet Model Visual Representation. Retrieved from [2]	43
Figure 2.22	Preprocessing Steps in Proposed Method in [3]. Adapted from [3]	44
Figure 2.23	Proposed Model Architecture in Paper [3]. Adapted from [3]	46
Figure 2.24	YAMNet Model for Bird Sounds Classification. Retrieved from [4]	47
Figure 2.25	Used RawNet Model Shows in Below Part. Adapted from [5]	48
Figure 2.26	Binary Feature Set Performance Proposed in [6] Adapted from [6]	49
Figure 2.27	HMM along with GMM State Observations Tailored to Match The Characteristics of An Asian Elephant Squeak Vocalization. Retrieved from [7]	52

Figure 2.28	Confusion Matrix for Elephant Caller Type Classification in [7]. Retrieved from [7]	53
Figure 3.1	Working Flow of Research Implementation	58
Figure 3.2	Methodology Flowchart for Elephant Call Type Recognition Using Raw Audio and Audio Spectrogram Processing	60
Figure 3.3	Data Pre-Processing	63
Figure 3.4	Wavelet Denoising for 03 Main Caller Types: Roar, Trumpet, And Rumble	70
Figure 3.5	Gaussian Denoising for 03 Main Caller Types: Roar, Trumpet, And Rumble	72
Figure 3.6	MobileNet Model Basic Architecture	78
Figure 3.7	Components of Each Block in MobileNet	80
Figure 3.8	Components of MobileNet V2 Model	80
Figure 3.9	Architecture of Experimented YAMNet Model	83
Figure 3.10	Architecture of Separable Convolution Layer in YAMNet Model	84
Figure 3.11	Graphical Representation of Experimented RawNet Model	85
Figure 3.12	Residual Block Component of RawNet Model	85
Figure 3.13	GRU Component of RawNet Model	86
Figure 3.14	Architecture of ACDNet-based ElephantCallerNet Model	87
Figure 3.15	Elephant Sound Identification via ResNet18 Model	94
Figure 3.16	Example Application of Classification Model	101
Figure 3.17	Prototype Application : GUI Interface	102
Figure 4.1	Performance Comparison Between Smaller and Larger Datasets for Raw Audio	104
Figure 4.2	Comparative Analysis of Accuracy for Raw Audio Between Smaller and Larger Datasets Using YAMNet, RawNet, MobileNetV2, and ElephantCallerNet	105
Figure 4.3	Confusion Matrix for YAMNet, RawNet, MobileNetV2, and Elephant-CallerNet Models Trained on Larger Dataset with Raw Audio Processing	106
Figure 4.4	Confusion Matrix for Larger Dataset with Spectrogram Based Training using MobileNetV2, YAMNet, ResNet18, and SVM	108
Figure 4.5	Bar Plot of Accuracy Scores for Elephant Call Type Classification Using Implemented Models for Raw Audio	111
Figure 4.6	Comparative Analysis of Model Accuracy, Parameter Count, and Size across Implemented Deep Learning Models for Raw Audio	112
Figure 4.7	Radar Chart Illustrating Performance Metrics Across Implemented Deep Learning Models for Elephant Call Type Classification Using Raw Audio: Precision, Recall, and F1-score	114
Figure 4.8	Graph Depicting Inference Times for Implemented Deep Learning Models in Elephant Call Type Classification Using Raw Audio	117

Figure 4.9	Confusion Matrix for MobileNetV2 Model in Elephant Call Type Classification for Raw Audio Processing: Roar, Rumble, Trumpet	118
Figure 4.10	Training and Validation Loss and Accuracy Curves of MobileNetV2 Model for Raw Audio Processing	118
Figure 4.11	Confusion Matrix of YAMNet Model in Elephant Call Type Classification for Raw Audio Processing: Roar, Rumble, Trumpet	120
Figure 4.12	YAMNet Model Training Validation Accuracy and Loss Curves for Raw Audio Processing	122
Figure 4.13	RawNet Model Training Validation Accuracy and Loss Curves for Raw Audio Processing	122
Figure 4.14	Confusion Matrix of RawNet Model in Elephant Call Type Classification for Raw Audio Processing: Roar, Rumble, Trumpet	123
Figure 4.15	Confusion Matrix of ElephantCallerNet Model in Elephant Call Type Classification for Raw Audio Processing: Roar, Rumble, Trumpet	125
Figure 4.16	Training and Validation Loss and Accuracy Curves of ElephantCallerNet Model for Raw Audio Processing	125
Figure 4.17	Confusion Matrix for Spectrogram Analysis for Each Caller Type of Test Dataset using MobileNetv2, SVM, ResNet18, and YAMNet	127
Figure 4.18	Each Implemented Models' Validation and Training Curves for Spectrogram Analysing Using ResNet18, MobilnetV2, and YAMNet	129
Figure 4.19	Validation and Training, Loss and Accuracy Curves for Spectrogram Analysing Using ResNet18, MobilnetV2, and YAMNet	130
Figure 4.20	Comparative Analysis of Raw Audio Processing Vs Audio Spectrogram Processing for Experimented Results Using Smaller Dataset and Literature Review	132

LIST OF TABLES

Table	Description	Page
Table 2.1	Features Of LDC2010S05 Asian Elephant Vocalization Dataset	13
Table 2.2	Common Audio Feature Extraction Algorithms	25
Table 2.3	Summary of Elephant Calling Activity Detection in Previous Studies	32
Table 2.4	Comparison of Literature for High Accuracy Studies	33
Table 2.5	Table of Performance Metrics Analyses of the Model. Adapted from (Ranasinghe et al., 2023)	45
Table 3.1	Data Collection Summary	64
Table 3.2	Train and Evaluation Data Split Before Data Augmentation	66
Table 3.3	Data Split For Training And Evaluation After Augmentation to Create a Smaller Dataset	68
Table 3.4	Data Split For Training And Evaluation After Augmentation to Create a Larger Dataset	69
Table 3.5	MSE, SNR Values For Two Noise Filtering Techniques	71
Table 3.6	Number Of Features Extracted From Each Feature Type	76
Table 4.1	Comparative Analysis of Model Accuracy with Audio-Visual Training (Using MFCC & Chroma_cqt Features) for Smaller and Larger Datasets	107
Table 4.2	Performance Metrics for Larger Dataset with Spectrogram-Based Training using MobileNetV2, YAMNet, ResNet18, and SVM	107
Table 4.3	Performance Comparison of Implemented Models for Elephant Call Type Recognition Using Raw Audio	110
Table 4.4	Classification Accuracy Scores for Different Elephant Call Types Across Implemented Models for Raw Audio	111
Table 4.5	Comparative Analysis of Implemented Model's Accuracy and Parameter Count for Elephant Call Type Classification using Raw Audio	113
Table 4.6	Comprehensive Performance Evaluation of Implemented Deep Learning Models for Elephant Call Type Classification Using Raw Audio: Accuracy, Precision, Recall, F1-score, and Inference Time	115
Table 4.7	MobileNetV2 Model training Hyper-parameters for Raw Audio Processing	119
Table 4.8	Best Validation Accuracy of MobileNetV2 With Different Hyper-parameters for Raw Audio Processing	119
Table 4.9	YAMNet Model training Hyper-parameters for Raw Audio Processing	121
Table 4.10	Best Validation Accuracy of YAMNet With Different Hyper-parameters for Raw Audio Processing	121

Table 4.11	RawNet Model training Hyper-parameters for Raw Audio Processing	123
Table 4.12	Best Validation Accuracy of RawNet With Different Hyper-parameters for Raw Audio Processing	124
Table 4.13	ElephantCallerNet Model training Hyper-parameters for Raw Audio Processing	124
Table 4.14	Comparative Analysis of Model Accuracy With Audio-Visual Training (Using MFCC & Chroma_cqt Features)	131
Table 4.15	Inference Time, Parameter Count and Accuracy Comparison Between Raw Audio Processing and Audio Spectrogram Processing for Smaller Dataset	133
Table 5.1	Experimented Results Comparison With Literature Review For Full Test Dataset Using Smaller Dataset	152
Table 5.2	Experimented Results Comparison With Literature Review For Rumble Only Using Smaller Dataset	153

LIST OF ABBREVIATIONS

Abbreviation	Description
CNN	convolutional neural networks
DFT	Discrete Fourier Transform
ECN	ElephantCallerNet
FFT	Fast Fourier Transform
gPLP	generalized Perceptual Linear Prediction
HMM	Hidden Markov Model
HNR	Harmonics-to-Noise Ratio
LPC	Linear Predictive Coding
LSTM	Long Short-Term Memory
MCC	Matthews Correlation Coefficient
MFCCs	Mel-Frequency Cepstral Coefficients
SFEB	Spectral Feature Extraction Block
SSCs	Spectral Subband Centroids
STFT	short-time Fourier transform
SVM	Support Vector Machine
TFEB	Temporal Feature Extraction Block
YAMNet	<i>Yet Another Mobile Network</i>
ZCR	Zero-Crossing Rate

LIST OF APPENDICES

Appendix	Description	Page
Appendix -A	Accoustic Features for Each Elephant Caller Type	165
Appendix -B	Code Bases	171