

LLB/TH/43/2025
TH6009

**INTEGRATING MUSIC INFORMATION RETRIEVAL
AND TRANSFER LEARNING FOR ADVANCED
EMOTION RECOGNITION IN SRI LANKAN CROWD
SOUNDSCAPES: DATASET CREATION AND
ANALYSIS**

P.B.S.N.Ariyathilake

219311M

MSc in Computer Science

Department of Computer Science & Engineering
Faculty of Engineering

University of Moratuwa
Sri Lanka

June 2025

**INTEGRATING MUSIC INFORMATION RETRIEVAL
AND TRANSFER LEARNING FOR ADVANCED
EMOTION RECOGNITION IN SRI LANKAN CROWD
SOUNDSCAPES: DATASET CREATION AND
ANALYSIS**

P.B.S.N.Ariyathilake

219311M

Thesis/Dissertation submitted in partial fulfillment of the requirements for the degree
MSc in Computer Science

Department of Computer Science & Engineering
Faculty of Engineering

University of Moratuwa
Sri Lanka

June 2025

DECLARATION

I declare that this is my work and this thesis/dissertation 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: *UOM Verified Signature*

Date: 18.06.2025

The above candidate has carried out research for the PhD/MPhil/Master's thesis/dissertation under my supervision. I confirm that the declaration made above by the student is true and correct.

Name of Supervisor: Dr. Charith Chithranjan

Signature of the Supervisor:

Date: 18.06.2025

ACKNOWLEDGEMENT

First and foremost, I am deeply grateful to the University of Moratuwa and the Department of Computer Science & Engineering for providing me with the opportunity and resources to pursue my Master's studies.

I am deeply grateful to my supervisor, Dr. Charith Chitraranjan, for his continuous encouragement, guidance, understanding, and support throughout my research journey. His invaluable advice, constructive feedback, and inspiration greatly helped me to grow both theoretically and practically in this subject area. I am forever grateful for his mentorship.

I also extend my heartfelt thanks to Dr. Chathuranga Hettiarachchi, the MSc Coordinator, for his insightful discussions, advice, and support, which played an important role in the successful completion of this research.

A special word of appreciation goes to all the lecturers who taught during the MSc program. I am truly thankful for their dedication, time, and the invaluable knowledge they imparted.

I am also grateful to my friend Tharika and Shyamalee, whose willingness to share their own experiences helped me stay motivated and focused throughout this thesis journey.

I am deeply indebted to my parents, especially my mother, for constantly encouraging me to complete this research and for the unwavering support she has given, and to my sister for always checking on my progress, which motivated me to stay focused.

Finally, I would like to extend a heartfelt thank you to my husband for being a source of motivation and helping me in every means, where he can. And to my baby boy for the sacrifices made in sharing time with me during this journey. Their love and patience were my strength throughout.

ABSTRACT

Understanding crowd emotions through sound is critical for applications in event monitoring, public safety, and mental health studies. However, there has been a notable gap in the availability of specialized datasets and novel robust models for classifying crowd sound emotions. To address this, a comprehensive Sri Lankan Crowd Sound Emotion Dataset (SLCSED) was developed, enriched with detailed annotations, to support future research. The study proposes a computational framework based on Music Information Retrieval (MIR) techniques combined with advanced machine learning algorithms to perform emotion classification in crowd. Feature extraction was performed using MIR methods, Wav2Vec 2.0 embeddings, and Emotion2Vec representation. PCA was applied as a dimensionality reduction technique. Various machine learning and transfer learning classifiers, including TabNet, LightGBM, Multi-Layer Perceptrons (MLP), wav2vec, and emotion2vec, were evaluated. Specific architectures were tuned for better accuracy, such as LightGBM with Gradient boosting and MLPs with hidden layers of (128, 64) units. Furthermore, emotion recognition models were developed using supervised learning methods, drawing inspiration from approaches tested on decision trees, random forests, XGBoost, and LightGBM in related studies. The results demonstrated highly promising outcomes, with the LightGBM classifier achieving up to 99.95% validation accuracy on the Emotional Crowd Sounds Data (ECSD) dataset and the MLP achieving 99.53% on the SLCSED dataset without dimensionality reduction. PCA was found to slightly reduce the performance in most cases. Additionally, the Emotion2Vec framework showed significant improvements after PCA application, reaching 99.99% accuracy. These findings highlight the effectiveness of MIR-based feature engineering combined with carefully selected classifiers for crowd emotion detection. This work not only fills a major gap by introducing a localized and richly annotated dataset but also presents a robust methodological pipeline for crowd sound emotion recognition, paving the way for future applications in real-world monitoring and psychological analysis.

Keywords: ECSD, MIR, MLP, PCA, SLCSED

TABLE OF CONTENTS

Declaration	i
Acknowledgement.....	ii
Abstract	iii
Table of Contents	iv
List of Figures	ix
List of Tables.....	xii
List of Abbreviations.....	xiv
Chapter 1	1
Introduction	1
1.1 Overview of the research.....	1
1.2 Importance of the research	3
1.3 Motivation for the research	5
1.4 Problem Statement.....	5
1.5 Research Objectives	7
1.6 Research Outcomes	7
Chapter 2	8
Literature Review	8
2.1 Overview of the Crowd Sound Emotions	8
2.2 The Role/Applications of Sound in Emotion Recognition	8
2.3 Studies in Crowd Sound Emotion Identification	9
2.3.1 Types of Crowd Emotions	10
2.3.2 Datasets Used for Crowd Sound Emotion Identification.....	10
2.3.3 Feature extraction of crowd sound emotions	11
2.3.4 Machine Learning/Deep Learning models in crowd sound emotions	11
2.3.5 Transfer Learning models in crowd sound emotions.....	12
2.4 Speech Emotion Recognition.....	13
2.5 TabNet in Speech Emotion Recognition.....	13
2.6 LightGBM in Speech Emotion Recognition	14
2.7 Multi-Layer Perceptron in Speech Emotion Recognition.....	14
2.8 Wav2vec in Speech Emotion Identification.....	15
2.9 emotion2vec in Speech Emotion Identification	15
2.10 Music Information Retrieval in Music	15
2.11 Applicability of frequency domain features extracted using the Fourier transform for SER	16

Chapter 3	18
Methodology	18
3.1 Data for the research.....	18
3.1.1 Sri Lankan context crowd emotion audio dataset preparation	18
3.1.2 Sri Lankan context crowd emotion audio dataset validation	25
3.1.3 Emotional Crowd Sounds dataset(ECSD) from the IEEE data portal[74]26	
3.2 Classification Approach	27
3.2.1 Overall process.....	27
3.2.1 Experiment 01	29
3.2.2 Experiment 02	35
3.2.3 Experiment 03	37
3.2.4 Experiment 04	38
3.2.5 Experiment 05	38
3.2.6 Experiment 06	39
3.2.7 Experiment 07	39
3.2.8 Experiment 08	41
3.2.9 Experiment 09	41
3.2.10 Experiment 10	42
3.2.11 Experiment 11	42
3.2.12 Experiment 12	43
3.2.13 Experiment 13	44
3.2.14 Experiment 14	44
3.2.15 Experiment 15	44
3.2.16 Experiment 16	45
3.2.17 Experiment 17	45
3.2.18 Experiment 18	45
3.3 SLCSSED with Existing Algorithms	47
3.3.1 Experiment 19	47
3.3.2 Experiment 20	48
3.3.3 Experiment 21	48
3.3.4 Experiment 22	49
3.3.5 Experiment 23	49
3.3.6 Experiment 24	50
3.3.7 Experiment 25	50
3.4 Applicability of frequency domain features extracted using the Fourier transform for Crow Sound emotion data.....	51
3.4.1 Experiment 26	54

3.4.2 Experiment 27	55
3.4.3 Experiment 28	55
Chapter 4	57
Results and Discussion.....	57
4.1 Results Comparison	57
4.1. Results grouped by feature extraction technique	57
4.2 Results grouped by PCA usage	59
4.3 Results grouped by Classifier.....	61
4.4 Results grouped by dataset.....	63
4.5 Combined View Accuracy by Classifier with and without PCA.....	65
4.6 Combined View: Feature Extraction × Classifier × PCA	66
4.7 Results Experiment-wise.....	67
4.2 Overall Discussion	68
4.3 Results comparison with existing works.....	69
4.4 SLCSSED dataset performance: with existing classification algorithms	73
4.5 Applicability of frequency domain features extracted using the Fourier transform on crowd sound emotion data.....	75
Chapter 5	77
Conclusion.....	77
5.1 Conclusion.....	77
5.2 Sri Lankan Crowd Sound Emotion Dataset - SLCSSED.....	79
Reference.....	80
Appendix A	93
Experiment 01: Plots and Graphs.....	93
Appendix B	96
Experiment 02: Plots and Graphs.....	96
Appendix C	102
Experiment 03: Plots and Graphs.....	102
Appendix D	105
Experiment 04: Plots and Graphs.....	105
Appendix E.....	108
Experiment 05: Plots and Graphs.....	108
Appendix F	111
Experiment 06: Plots and Graphs.....	111
Appendix G	114
Experiment 07: Plots and Graphs.....	114
Appendix H	116

Experiment 08: Plots and Graphs.....	116
Appendix I.....	118
Experiment 09: Plots and Graphs.....	118
Appendix J.....	119
Experiment 10: Plots and Graphs.....	119
Appendix K.....	120
Experiment 11: Plots and Graphs.....	120
Appendix L.....	123
Experiment 12: Plots and Graphs.....	123
Appendix M.....	125
Experiment 13: Plots and Graphs.....	125
Appendix N.....	128
Experiment 14: Plots and Graphs.....	128
Appendix O.....	131
Experiment 15: Plots and Graphs.....	131
Appendix P.....	134
Experiment 16: Plots and Graphs.....	134
Appendix Q.....	137
Experiment 17: Plots and Graphs.....	137
Appendix R.....	139
Experiment 18: Plots and Graphs.....	139
Appendix S.....	142
Experiment 19: Plots and Graphs.....	142
Appendix T.....	143
Experiment 20: Plots and Graphs.....	143
Appendix U.....	145
Experiment 21: Plots and Graphs.....	145
Appendix V.....	147
Experiment 22: Plots and Graphs.....	147
Appendix W.....	149
Experiment 23: Plots and Graphs.....	149
Appendix X.....	151
Experiment 24: Plots and Graphs.....	151
Appendix Y.....	153
Experiment 25: Plots and Graphs.....	153
Appendix Z.....	154
Experiment 26: Plots and Graphs.....	154

Appendix AA	156
Experiment 27: Plots and Graphs.....	156
Appendix AB	158
Experiment 28: Plots and Graphs.....	158

LIST OF FIGURES

Figure	Description	Page
Fig. 3.1:	Normalization process	21
Fig. 3.2:	Manual Revision Process	23
Fig. 3.3:	Manual Revision Process on terminal	23
Fig. 3.4 :	Complete Dataset Creation Process	24
Fig. 3.5:	Dataset Validation process	26
Fig. 3.6:	Overall classification Methodology	27
Fig. 3.7:	MIR architecture	31
Fig. 3.8:	Architecture of used model	34
Fig. 3.9:	PCA workflow	37
Fig. 3.10:	Wav2vec model architecture Adapted from [70]	40
Fig. 3.11:	Model Architecture	43
Fig. 3.12:	SLCSED Sub-methodology	47
Fig. 3.13:	Waveform of a selected audio	51
Fig. 3.14:	FFT Magnitude Spectrum	52
Fig. 3.15:	Log Power Spectrum	52
Fig. 3.16:	Spectrograms of the dataset	52
Fig. 3.17:	Average FFT Spectrum per emotion	53
Fig. 3.18:	Overall sub-methodology with FFT	54
Fig. 4.1:	Feature Extraction Technique Comparison Box Plot	58
Fig. 4.2:	Effect of PCA Comparison Box Plot	60
Fig. 4.3:	Classifier Comparison Box Plot	62
Fig. 4.4:	Accuracy of Dataset Comparison Box Plot	64
Fig. 4.5:	Accuracy by Classifier with and without PCA	65
Fig. 4.6:	Combined View: Feature Extraction × Classifier × PCA	66
Fig. 4.7:	Accuracies of the Experiments Comparison	68
Fig. A.1:	Validation Accuracy over Folds	93
Fig. A.2:	Training Loss over Folds	94
Fig. A.3:	Confusion Matrix	94
Fig. A.4:	ROC Curve	95
Fig. A.5:	Classification Report Heatmap	95
Fig. B.1	Validation Accuracy across the 5 folds	96
Fig. B.2:	Training loss over folds	97
Fig. B.3	Confusion Matrix	97
Fig. B.4	ROC Curve	98
Fig. B.5	Classification Report Heatmap	98
Fig. B.6	PCA Explained variance	99
Fig. B.7	Feature contribution	99
Fig. B.8	2D Plot	100
Fig. B.9	3D Plot	100
Fig. B.10	Feature Importance from TabNet	101
Fig. C.1	Validation Accuracy.	102
Fig. C.2	Training loss vs validation loss	103
Fig. C.3	Confusion Matrix	103
Fig. C.4	ROC Curve	104
Fig. C.5	Classification report heatmap	104

Fig. D.1 Validation Accuracy	105
Fig. D.2 Training loss vs validation loss	106
Fig. D.3 Confusion Matrix	106
Fig. D.4 ROC Curve	107
Fig. D.5 Classification Report Heatmap	107
Fig. E.1: Validation Accuracy over Folds	108
Fig. E.2: Training Loss across Folds	109
Fig. E.3: Confusion Matrix	109
Fig. E.4: ROC Curve	110
Fig. F.1: Validation Accuracy over the Folds	111
Fig. F.2: Training loss over the Folds	112
Fig. F.3: Confusion Matrix	112
Fig. F.4: ROC Curve	113
Fig. G.1: Validation Accuracy over Folds	114
Fig. G.2: Training loss over the Folds	115
Fig. G.3: Confusion Matrix	115
Fig. H.1: Validation Accuracy over Epochs	116
Fig. H.2: Training loss over the Epochs	117
Fig. H.3: Confusion Matrix	117
Fig. I.1: Confusion Matrix	118
Fig. J.1: Confusion Matrix	119
Fig. K.1: Validation accuracy	120
Fig. K.2: Training Accuracy across folds	120
Fig. K.3: Validation loss	121
Fig. K.4: Confusion Matrix	122
Fig. L.1: Validation accuracy	123
Fig. L.2: Training Accuracy	123
Fig. L.3: Validation loss	124
Fig. L.4: Confusion Matrix	124
Fig. M.1: Validation accuracy over folds	125
Fig. M.2: Training vs Validation loss	126
Fig. M.3: Confusion Matrix	126
Fig. M.4: ROC Curve	127
Fig. N.1: Validation accuracy over folds	128
Fig. N.2: Training vs Validation loss	129
Fig. N.3: Confusion Matrix	129
Fig. N.4: ROC Curve	130
Fig. O.1: Validation accuracy over folds	131
Fig. O.2: Training accuracy over folds	132
Fig. O.3: Validation loss	132
Fig. O.4: Confusion Matrix	133
Fig. P.1: Validation accuracy over folds	134
Fig. P.2: Training accuracy over folds	135
Fig. P.3: Validation loss	135
Fig. P.4: Confusion Matrix	136
Fig. Q.1: Validation accuracy over Folds	137
Fig. Q.2: Training loss across Folds	137
Fig. Q.3: ROC Curve	138
Fig. R.1: Validation accuracy over folds	139
Fig. R.2: Training loss across Folds	140

Fig. R.3:Confusion Matrix	140
Fig. R.4:ROC Curve	141
Fig. S.1: Confusion Matrix	142
Fig. T.1:Validation accuracy over folds	143
Fig. T.2: Confusion Matrix	144
Fig. T.3: ROC Curve	144
Fig. U.1:Validation accuracy over folds	145
Fig. U.2: Confusion Matrix	146
Fig. U.3: ROC Curve	146
Fig. V.1:Validation accuracy over folds	147
Fig. V.2: Confusion Matrix	148
Fig. V.3: ROC Curve	148
Fig. W.1:Validation accuracy over folds	149
Fig. W.2: Confusion Matrix	150
Fig. W.3: ROC Curve	150
Fig. X.1:Validation accuracy over folds	151
Fig. X.2: Confusion Matrix	152
Fig. Y.1: Confusion Matrix	153
Fig. Z.1:Validation accuracy over folds	154
Fig. Z.2:Confusion Matrix	155
Fig. Z.3 :ROC Curve	155
Fig. AA.1:Validation accuracy over folds	156
Fig. AA.2: Confusion Matrix	157
Fig. AA.3: ROC Curve	157
Fig. AB.1:Validation accuracy over folds	158
Fig. AB.2:Confusion Matrix	159

LIST OF TABLES

Table	Description	Page
TABLE 2.1:	ECSD Dataset Description	11
TABLE 2.2:	Summary of the Existing Literature	12
TABLE 2.3:	MIR Applications	16
TABLE 3.1:	Original Audio Set Description	19
TABLE 3.2:	Number of Audio Files Collected	19
TABLE 3.3:	Number of Audio Segments	22
TABLE 3.4:	Number of Audio Removed	23
TABLE 3.5:	Number of Audio Remained After Removal	24
TABLE 3.6:	Number Of Audio Remained After Validation	25
TABLE 3.7:	Dataset Description ECSD	26
TABLE 3.8:	Experiment Summary	28
TABLE 3.9:	MIR Features	30
TABLE 3.10:	Hyperparameters of Model	33
TABLE 3.11:	Hyperparameters of The Model	35
TABLE 3.12:	Hyperparameters of The LightGBM	38
TABLE 3.13:	Model architecture	39
TABLE 3.14:	Hyperparameters of MLP	39
TABLE 3.15:	Hyperparameters of MLP	40
TABLE 3.16:	Hyperparameters of MLP	41
TABLE 3.17:	LightGBM Architecture	41
TABLE 3.18:	LightGBM Architecture	42
TABLE 3.19:	Emotion2vec Architecture	43
TABLE 3.20:	Emotion2vec Architecture	44
TABLE 3.21:	LightGBM Architecture	44
TABLE 3.22:	LightGBM Architecture	44
TABLE 3.23:	Emotion2vec Architecture	45
TABLE 3.24:	Emotion2vec Architecture	45
TABLE 3.25:	Hyperparameters of The MLP Model	45
TABLE 3.26:	Hyperparameters of The MLP Model	45
TABLE 3.27:	Hyperparameters of The AlexNet	48
TABLE 3.28:	Hyperparameters of The SVM Model	48
TABLE 3.29:	Hyperparameters of The RF Model	48
TABLE 3.30:	Hyperparameters of The KNN Model	49
TABLE 3.31:	Hyperparameters of The ID CNN Model	49
TABLE 3.32:	Hyperparameters of The 2D CNN Model	50
TABLE 3.33:	Hyperparameters of The Vision Transformer(ViT) Model	51
TABLE 3.34:	Hyperparameters of The LightGBM Model	54
TABLE 3.35:	Hyperparameters of The MLP Model	55
TABLE 3.36:	Hyperparameters of The TabNet Model	55
TABLE 4.1 :	MIR Technique Wise	57
TABLE 4.2 :	Wav2vec 2.0 Wise	57
TABLE 4.3 :	Emotion2vec 2.0 Wise	58
TABLE 4.4 :	Without PCA- Performance	59
TABLE 4.5 :	With PCA- Performance	59
TABLE 4.6 :	TabNet Wise	61

TABLE 4.7 : LightGBM Wise	61
TABLE 4.8 : MLP wise	61
TABLE 4.9 : Emotion2vec Wise	62
TABLE 4.10 : SLCSED-wise	63
TABLE 4.11 : ECSD-wise	63
TABLE 4.12 : Results Experiment Wise	67
TABLE 4.13 : Architecture Of Some Experiments	69
TABLE 4.14: Comparison of SLCSED Dataset with existing datasets	70
TABLE 4.15: Performance comparison with Existing Literature	71
TABLE 4.16 : Performance Comparison with ECSD Dataset	72
TABLE 4.17: SLCSED dataset performance: with existing classification algorithms	73
TABLE 4.18: Performance With FFT Algorithm	75
TABLE 4.19: MIR Vs FFT as Feature Extraction	75
TABLE A.1 : Precision, Recall, F1 Score And Support	93
TABLE B.1 : Precision, Recall, F1 Score And Support	96
TABLE C.1 : Precision, Recall , F1 Score And Support	102
TABLE D.1 : Precision, Recall, F1 Score And Support	105
TABLE E.1 : Precision, Recall, F1 Score And Support	108
TABLE F.1 : Precision, Recall, F1 Score And Support	111
TABLE G.1 : Precision, Recall, F1 Score And Support	114
TABLE H.1 : Precision, Recall, F1 Score And Support	116
TABLE K.1 : Precision, Recall, F1 Score And Support	121
TABLE L.1 : Precision, Recall, F1 Score And Support	124
TABLE M.1 : Precision, Recall, F1 Score And Support	125
TABLE N.1 : Precision, Recall, F1 Score And Support	128
TABLE O.1 : Precision, Recall, F1 Score And Support	131
TABLE P.1 : Precision, Recall, F1 Score And Support	134
TABLE Q.1 : Precision, Recall, F1 Score And Support	137
TABLE R.1 : Precision, Recall, F1 Score And Support	139
TABLE S.1 : Precision, Recall, F1 Score And Support	142
TABLE T.1 : Precision, Recall, F1 Score And Support	143
TABLE U.1 : Precision, Recall, F1 Score And Support	145
TABLE V.1 : Precision, Recall, F1 Score And Support	147
TABLE W.1 : Precision, Recall, F1 Score And Support	149
TABLE X.1 : Precision, Recall, F1 Score And Support	151
TABLE Y.1 : Precision, Recall, F1 Score And Support	153
TABLE Z.1 : Precision, Recall, F1 Score And Support	154
TABLE AA.1 : Precision, Recall, F1 Score And Support	156
TABLE AB.1 : Precision, Recall, F1 Score And Support	158

LIST OF ABBREVIATIONS

Abbreviation	Description
CNN	Convolutional Neural Networks
CSV	Comma Separated Values
CTC	Connectionist Temporal Classification
ECSD	Emotional Crowd Sound Data
FFT	Fast Fourier Transform
FN	False Negative
FP	False Positive
FP	Fourier parameter
GBDT	Gradient Boosting Decision Tree
GBM	Gradient Boosting Machine
KNN	K-Nearest Neighbors
LTD	Linear Discriminant Analysis
MFCC	Mel-frequency cepstral coefficients
MIR	Music Information Retrieval
MLi	Multi-Lingual
MLP	Multi-Layer Perceptrons
MTL	Multi-Task Learning
PANN	Pre-trained Audio Neural Network
PCA	Principal Component Analysis
PNN	Probabilistic Neural Network
RF	Random Forest
ROC	Receiver-operating characteristic
SER	Speech Emotion Recognition
SLCSED	Sri Lankan Crowd Sound Emotion Dataset
SVM	Support Vector Machine
TIM-NET	Temporal-aware bi-direction Multi-scale Network

TN	True Negative
TP	True Positive
ViT	Vision Transformer