ADScAI
Applied Data Science & Artificial Intelligence Symposium

Explainable AI for Speech Emotion Recognition

S.S.J. Patabendige
Department of Computer science and Engineering,
Faculty of Engineering
University of Moratuwa,
Katubedda, 10400 Sri Lanka
Susarajayaweera95@gmail.com

Dr Uthayasanker Thayasivam
Department of Computer science and Engineering,
Faculty of Engineering
University of Moratuwa,
Katubedda, 10400 Sri Lanka

Keywords— *Explainable AI (XAI), ANN, ML Models, SER*

I. ABSTRACT

Artificial Intelligence (AI) has become essential across domains, excelling in classification, regression, clustering, and optimization [1]. However, the opacity of traditional AI models, particularly in Speech Emotion Recognition (SER), highlights the need for greater explainability [1]. This research advances Explainable AI (XAI) by developing SER models [2], [3]. It integrates insights from a Literature Review, enhances human-centered XAI methods, and utilizes 18 features for analysis. A feature range metric assesses model performance and explanation quality [4], contributing to a more transparent and interpretable AI framework for SER.

A. Introduction

Artificial Intelligence (AI) has increasingly influenced various domains, including healthcare, defense, and governance [1], [5], [6]. However, the opacity of AI models, particularly in Speech Emotion Recognition (SER), raises concerns about accountability and user trust [1]. Traditional AI functions as a black box, often failing to explain its decisions, especially regarding crucial "wh" questions like why, when, and where [7].

Explainable AI (XAI) addresses these challenges by enhancing transparency and interpretability in AI systems [1], [4]. This research focuses on SER, emphasizing the need to understand AI-driven decisions [8]. Using Artificial Neural Networks (ANN), Random Forest, Support Vector Machine (SVM), XGBoost, and k-Nearest Neighbors (k-NN), we aim to improve explainability in SER [2], [3]. The need for this research arises from the limitations of conventional AI models in explaining their decision-making processes [7]. As voice assistants like Alexa and Siri rely on audio waveform-based classification, enhancing XAI in SER is increasingly important [1]. Virtual agents further highlight the necessity of explainability and visual representation to foster user trust [1], [5]. This study follows a structured methodology, beginning with a comprehensive Literature Review on XAI in SER [9]. It then focuses on improving human-centered XAI methods, designing interpretable models, defining evaluation metrics, and documenting findings [2], [3]. By integrating explainability into AI-driven SER, this research aims to enhance transparency and accountability in voice-based applications using a value matrix [8].

B. Literature Review

Explainable Artificial Intelligence (XAI) enhances the interpretability of Speech Emotion Recognition (SER) systems. SHAP (SHapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations) provide transparent insights into model decisions [9]. SHAP's foundation was established by Lundberg and Lee [9], while Ribeiro et al. introduced LIME for local interpretability [10]. These techniques build on Ekman's work on emotional expressions [11] and support various SER models, including Han et al.'s Explainable Enhanced Recurrent Neural Network (ERNN), which optimizes LSTM hyperparameters using fuzzy logic [4]. Research explores diverse feature sets, such as MFCC (achieving 98.75%-100% accuracy in isolated speech recognition [12]) and multimodal audio-visual approaches [4]. By addressing the black-box nature of models like bidirectional LSTMs [13], XAI fosters trust and improves SER applications, from stress detection [4] to deception analysis [14].

B. Materials and Methods

The methodology followed four sequential steps. First, a comprehensive literature review established the state of XAI in voice embedding for speech emotion recognition, identifying research gaps. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) was used, comprising 1440 recordings from 24 actors expressing seven emotions at varying intensities (calm, happy, sad, angry, fearful, surprise, and disgust). This academically certified dataset (Livingstone & Russo, 2018) ensured research reliability.

Next, feature extraction involved systematically deriving 18 acoustic features, including temporal, spectral, and perceptual characteristics. Key features—loudness, zero-crossing rate, tempo, speaking rate, pitch, spectral contrast, MFCCs, chroma, harmonic/percussive components, spectral rolloff, and log mel spectrograms—were selected for comprehensive signal representation, robustness, and relevance to emotion/speaker identification.

The third phase trained multiple machine learning models: ANN, Random Forest, SVM, XGBoost, and k-NN. Preprocessing included MinMaxScaler normalization, an 80:20 stratified train-test split, and SMOTE for class balance. Hyperparameter tuning was performed using GridSearchCV with 5-fold cross-validation to optimize generalization.

Finally, explainability was incorporated using LIME and SHAP. Feature importance was calculated for test-set samples, aggregated to avoid bias, and analyzed through statistical measures (minima, maxima, means, and standard deviations). This framework identifies emotion-specific acoustic signatures, enhancing interpretability. The study demonstrates how explainability strengthens voice-based emotion recognition, fostering trust for applications in healthcare, customer service, and human-computer interaction.

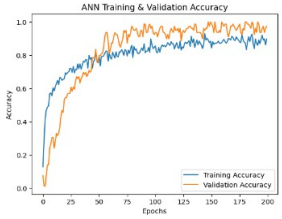
https://doi.org/10.31705/ADScAI.2025.30

C. *Results and Discussion*

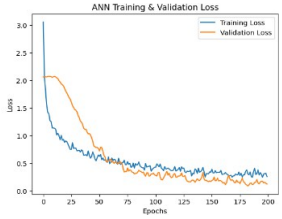
Below tables and graph are related to selected models and given results with model accuracy and XAI finalized results

Model	Best Parameters	Best CV Accuracy	Tuning Method
Random Forest	{'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 200}	0.7937	GridSearchCV
SVM	{'C': 100, 'gamma': 1, 'kernel': 'rbf'}	0.8318	GridSearchCV
XGBoost	{'colsample_bytree': 1.0, 'learning_rate': 0.1, 'max_depth': 6, 'n_estimators': 200, 'subsample': 0.8}	0.7963	GridSearchCV
k-NN	{'n_neighbors': 3, 'p': 1, 'weights': 'distance'}	0.6761	GridSearchCV
ANN	{'batch_size': 32, 'epochs': 200, 'optimizer': 'adam', 'dropout_rate': 0.3}	0.9620	GridSearchCV

Table 1: Selected model with best parameters



Graph 1: Accuracy



Graph 2: Loss

Class	Precision	Recall	F1-Score
Angry	0.61	0.92	0.73
Calm	1	0.75	0.86
Disgust	0.5	0.36	0.42
Fearful	0.71	0.62	0.67
Happy	0.33	0.22	0.27
Neutral	0.8	0.67	0.73
Sad	0.5	1	0.67
Surprised	0.75	0.75	0.75

Table 2: Random Forest

Class	Precision	Recall	F1-Score
Angry	1	0.92	0.96
Calm	0.89	0.67	0.76
Disgust	0.75	0.55	0.63
Fearful	0.71	0.62	0.67
Happy	0.64	0.78	0.7
Neutral	0.67	1	0.8
Sad	0.71	1	0.83
Surprised	0.85	0.92	0.88

Table 3: SVM

Class	Precision	Recall	F1-Score
Angry	0.69	0.92	0.79
Calm	1	0.67	0.8
Disgust	0.64	0.82	0.72
Fearful	0.71	0.62	0.67
Happy	0.5	0.22	0.31
Neutral	0.71	0.83	0.77
Sad	0.71	1	0.83
Surprised	0.75	0.75	0.75

Table 4: XG Boost

Class	Precision	Recall	F1-Score
Angry	0.92	0.92	0.92
Calm	0.86	0.5	0.63
Disgust	0.44	0.36	0.4
Fearful	0.67	0.5	0.57
Happy	0.15	0.22	0.18
Neutral	0.4	0.67	0.5
Sad	0.36	0.8	0.5
Surprised	0.57	0.33	0.42

Table 5: k-NN

Class	Precision	Recall	F1-Score
Angry	1	0.83	0.91
Calm	0.92	0.92	0.92
Disgust	0.71	0.45	0.56
Fearful	0.8	1	0.89
Happy	0.62	0.89	0.73
Neutral	1	0.83	0.91
Sad	0.8	0.8	0.8
Surprised	0.77	0.83	0.8

Table 6: ANN

Emotion	Feature	Min	Max	Mean	Std
Angry	mfcc_mean	-40.268	-26.4461	-34.3805	4.2676
	mfcc_std	27.4386	32.5181	30.1076	1.3455
	log_mel_spec_mean	-68.2463	-60.407	-65.5612	1.8104
	loudness	0.0042	0.058	0.031	0.016
	zcr	0.0521	0.0927	0.0706	0.0115
Calm	mfcc_mean	-61.3888	-38.7983	-48.8274	6.4354
	mfcc_std	24.9792	30.0357	27.738	1.2779
	loudness	0.0005	0.0078	0.0022	0.0017
	zcr	0.0409	0.1296	0.0592	0.0212
Disgust	mfcc_std	28.1765	33.9059	30.8722	1.4505
	loudness	0.0031	0.0512	0.0171	0.0132
	average_pitch	2604.358	6703.208	5388.353	811.1303
	zcr	0.0599	0.0921	0.0735	0.0103
	chroma_std	0.2767	0.3567	0.3122	0.0229
Fearful	zcr	0.0598	0.1138	0.087	0.0151
	log_mel_spec_mean	-67.8642	-58.9344	-64.2692	2.4524
	loudness	0.0031	0.0512	0.0171	0.0132
	mfcc_std	26.0945	30.6851	28.7484	1.372
	mfcc_mean	-48.0664	-32.6778	-41.3104	4.3193
Happy	log_mel_spec_mean	-68.2504	-58.7817	-63.8844	2.4875
	mfcc_std	27.2366	31.8342	29.9046	1.4629
	zcr	0.0389	0.1082	0.064	0.0211
	mfcc_mean	-51.6278	-32.9075	-40.424	5.2889
	loudness	0.0035	0.0337	0.0122	0.0094
Neutral	mfcc_mean	-55.2103	-38.7488	-47.5811	6.1325
	log_mel_spec_mean	-67.5405	-64.6839	-65.9283	0.973
	chroma_std	0.2697	0.3602	0.3255	0.0366
	zcr	0.0482	0.0617	0.0545	0.005
	loudness	0.0014	0.0051	0.0028	0.0011
Sad	mfcc_std	25.1723	31.7792	28.2067	1.999
	zcr	0.0319	0.0842	0.0642	0.019
	mfcc_mean	-55.8218	-36.5694	-46.2204	5.413
	log_mel_spec_mean	-70.0778	-61.5713	-65.0064	2.4931
	log_mel_spec_std	14.4117	20.3796	16.6329	1.7456
Surprised	log_mel_spec_mean	-69.4568	-62.3138	-65.3049	1.839
	mfcc_std	25.6213	33.3562	30.6808	1.907
	zcr	0.0355	0.1002	0.0652	0.016
	mfcc_mean	-51.9761	-35.3363	-44.2413	4.2861
	chroma_std	0.281	0.3636	0.3281	0.0239

Table 7: Selected important features and value range

This study evaluated feature importance in speech emotion recognition by selecting **SVM** and **ANN** based on model accuracy, followed by applying **LIME** and **SHAP** to interpret their decision-making processes. The top-ranked features for each emotion were determined by aggregating results from both explainability techniques, as presented in **Table 7**, along with their corresponding value ranges. This analysis identified critical acoustic thresholds—such as pitch variance for *anger* and spectral rolloff for *happiness*—that consistently influenced emotion classification. The findings demonstrate that certain features (e.g., MFCCs for *sadness*, zero-crossing rate for *fear*) exhibit robust importance across models and explanation methods, providing a statistically grounded framework for interpretable SER systems

D. *Conclusion*

This research significantly advances Explainable Artificial Intelligence (XAI) in Speech Emotion Recognition by introducing the Value of Feature Range's Metric, which quantifies feature importance across multiple models (k-NN, Random Forest, SVM, XGBoost, and ANN) and employs both LIME and SHAP analysis techniques to bridge the gap between model performance and interpretability. Through comprehensive analysis of 18 acoustic features from an mp3 dataset containing seven emotional labels, the study establishes a transparent framework that enhances human understanding of AI decision-making processes and provides statistical context for model predictions, ultimately creating more trustworthy and interpretable emotion recognition systems.

REFERENCES

1. P. Gohel, P. Singh, and M. Mohanty, "Explainable AI: current status and future directions," pp. 1–16, 2021, [Online]. Available: http://arxiv.org/abs/2107.07045.
2. T. G. Fantaye, J. Yu, and T. T. Hailu, "Advanced convolutional neural network-based hybrid acoustic models for low-resource speech recognition," Computers, vol. 9, no. 2, pp. 1–27, 2020, doi: 10.3390/computers9020036.
3. Y. Chen, Y. Zhou, S. Zhu, and H. Xu, "Detecting offensive language in social media to protect adolescent online safety," Proc. - 2012 ASE/IEEE Int. Conf. Privacy, Secur. Risk Trust 2012 ASE/IEEE Int. Conf. Soc. Comput. Soc. 2012, pp. 71–80, 2012, doi: 10.1109/SocialCom-PASSAT.2012.55.
4. F. M. Talaat, "Explainable Enhanced Recurrent Neural Network for lie detection using voice stress analysis," Multimed. Tools Appl., no. 0123456789, 2023, doi: 10.1007/s11042-023-16769-w.
5. W. Zhang and B. Y. Lim, "Towards Relatable Explainable AI with the Perceptual Process," Conf. Hum. Factors Comput. Syst. - Proc., 2022, doi: 10.1145/3491102.3501826.
6. S. Y. Lim, D. K. Chae, and S. C. Lee, "Detecting Deepfake Voice Using Explainable Deep Learning Techniques," Appl. Sci., vol. 12, no. 8, 2022, doi: 10.3390/app12083926.
7. S. Bekenova and A. Bekenova, "Emotion recognition and classification based on audio data using AI," E3S Web Conf., vol. 420, p. 10040, 2023, doi: 10.1051/e3sconf/202342010040.
8. M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," Speech Commun., vol. 116, pp. 56–76, 2020, doi:
9. K. Zhang, Y. Zhang, and M. Wang, "A Unified Approach to Interpreting Model Predictions Scott," Nips, vol. 16, no. 3, pp. 426–430, 2012.
10. M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?' Explaining the Predictions of Any Classifier," NAACL-HLT 2016 - 2016 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. Proc. Demonstr. Sess., pp. 97–101, 2016, doi: 10.18653/v1/n16-3020.
11. P. Ekman, "Lie Catching and Microexpressions," in The Philosophy of Deception, 2009, pp. 118–136.
12. G. Dede and M. H. Sazli, "Speech recognition with artificial neural networks," Digit. Signal Process. A Rev. J., vol. 20, no. 3, pp. 763–768, 2010, doi: 10.1016/j.dsp.2009.10.004.
13. G. I. Winata, O. P. Kampman, and P. Fung, "Attention-Based LSTM for Psychological Stress Detection from Spoken Language Using Distant Supervision," ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc., vol. 2018-April, pp. 6204–6208, 2018, doi: 10.1109/ICASSP.2018.8461990.
14. N. Palena, L. Caso, and A. Vrij, "Detecting lies via a Theme-Selection strategy," Front. Psychol., vol. 9, no. JAN, pp. 1–8, 2019, doi: 10.3389/fpsyg.2018.02775.