Melodic Mind: Machine Learning-Based Emotion Recognition for Personalized Music Recommendation

K Nithin, Dorai Sai Charan, Vijay Kumar, Kushal Yadav, Dr Kavyasai Yaddanapudi
Department of Computer Science and Engineering
Amrita School of Computing, Bengaluru
Amrita Vishwa Vidyapeetham, India
kandinithin154@gmail.com, doraisaicharan09@gmail.com,
vijaygosu99@gmail.com, kushalchelluboina09@gmail.com, y_kavya@blr.amrita.edu

***Abstract*—Melodic Mind is an adaptive music recommendation system that utilizes speech emotion recognition in real time to provide individualized music recommendations based on the user's present emotional state. The speech emotion classification makes use of the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), where audio signals are analyzed using a pretrained Convolutional Neural Network (CNN14) from the Pretrained Audio Neural Networks (PANNs) for extracting robust acoustic features. An MLP classifier is then trained on them in order to classify emotions such as happiness, sadness, anger, fear, surprise, and neutral. At the same time, music emotional content is also predicted using the MER500 dataset by ensemble of classifiers such as XGBoost, Random Forest, Support Vector Machine, and Multi-Layer Perceptron models. Through the combination of these elements, the system aligns identified speech emotions with music of similar emotional content and presents a context-aware and adaptive recommendation stream. This multimodal solution increases user experience through the creation of an emotionally connected listening experience. The applicability of the system ranges from mental health assistance and music therapy to personalized entertainment, exhibiting the potential of merging state-of-the-art audio-based emotion sensing with machine learning-based music recommendation.**

***Index Terms*—speech emotion recognition, music emotion classification, PANNs CNN14, personalized music recommendation**

I. Introduction

Music has been a strong tool for emotional expression and psychological control throughout history and across cultures, impacting mental health. Artificial intelligence has introduced emotion-aware computing on the rise, which improves individualized user experience through unobtrusive modalities such as speech emotion recognition (SER) [19], [20]. SER is central to affective computing and human-computer interaction, enabling machines to sense and react to emotional signals in human dialogue naturally and intuitively [16], [18]. With the international music streaming market booming towards an estimated value of more than USD 100 billion by 2030, emotion-based music recommendation systems are needed more and more [14], [15]. Traditional music recommendation systems primarily rely on static metrics like user preferences, listening history, or genre-based filtering, often neglecting the user's current emotional context [11], [17]. Addressing this limitation, recent research has explored emotion-centric approaches by integrating modalities such as facial recognition [11], deep learning-based music emotion classifiers [12], [13], and context-aware personalization strategies [16], [20]. These efforts highlight the growing relevance of dynamic, emotion-based systems that can adapt to user states in real time. We introduce in this paper Melodic Mind, a new multimodal system bringing together real-time speech emotion recognition and music emotion classification for the purpose of providing emotionally meaningful music suggestions. Speech emotion recognition is based on the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) and relies on a pre-trained CNN14 model from the PANNs toolkit in order to abstract high-level audio features. These are subsequently classified by a Multi-Layer Perceptron (MLP) classifier to identify user emotions of happiness, sadness, anger, fear, surprise, and neutrality. At the same time, the module of music emotion classification uses the MER500 dataset based on an ensemble of XGBoost, Random Forest, Support Vector Machine, and MLP models. To handle class imbalance, the Synthetic Minority Oversampling Technique (SMOTE) is used [15], [17]. The key contributions of this paper include, using pretrained deep learning models to extract robust speech emotion features and classify them [19], using ensemble learning techniques for proper music emotion classification [12], [13] and combining the above models in an integrated, real-time recommendation system that associates identified speech emotions with music of matching emotional content [11], [16], [20]. In comparison with standard recommendation systems, Melodic Mind provides an emotionally attuned and context-sensitive listening experience with important implications across areas such as mental health care, music therapy, and customized entertainment. The rest of the paper is structured as follows: Section 2 is a review of related work on emotion recognition and music recommendation. Section 3 is an explanation of the datasets, feature extraction techniques, model structures, and system pipeline. Section 4 gives the results and discusses them. Lastly, Section 5 concludes the paper with important findings, limitations, and suggestions for further research.

II. Literature Survey

The area of affective computing has seen great advances in speech and music emotion recognition, with scientists using more advanced methodologies to interpret emotional states from audio cues. Early methods in speech emotion recognition (SER) have mainly relied on acoustic features like Mel-frequency cepstral coefficients (MFCCs), pitch contours, and energy-based features along with traditional machine learning classifiers like Support Vector Machines (SVMs) and Gaussian Mixture Models (GMMs) [1], [10]. Parallel research in music emotion recognition (MER) has also depended on hand-engineered audio features such as MFCCs, chroma, and spectral features processed by traditional machine learning algorithms [3], [8]. Both fields are confronted with similar challenges such as the intrinsic subjectivity of emotional annotation, cultural and individual differences in perceiving emotions, and constraints imposed by acted datasets that might not be representative of true emotional expressions [1], [8]. The systematic reviews by [1] and [8] thoroughly examine these classical methods, pointing out their inability to capture the subtle richness of emotional states, especially in real-world applications where audio signals tend to be noisy and highly variable.

Recent breakthroughs have given rise to a new era with deep learning architectures and multimodal fusion methods ruling the roost in both SER and MER research. In SER, convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have become the dominant methods, with [9] showing the power of 1D-DCNNs with extensive acoustic feature sets to achieve state-of-the-art results on benchmarking datasets. The MER domain has also adopted deep learning, with [4] offering a thorough overview of different architectures such as CNNs, RNNs, and their hybrid forms for end-to-end learning from raw audio signals. A pioneering work is provided by [6], who proposed Music Theory-Inspired Acoustic Representation (MTAR) for SER, which actually filled the gap between music and speech analysis by utilizing musical elements like dynamics and intervals. Multimodal methods have attracted significant attention, especially in MER, where audio signals are combined with physiological signals (EEG) and facial expressions, as shown by [2] on the DEAP dataset. The [7] survey of automatic speech recognition has useful findings which can be ported over for emotion recognition, especially for transformer models and transfer learning methods. [5] introduce a new probabilistic technique based on Gaussian Processes that provides better performance than standard SVMs in music genre classification and emotion recognition tasks, whereas [3] investigate regression-based methods as a possible direction for continuous Valence-Arousal space emotion prediction.

The current state of speech and music emotion recognition shows both remarkable progress and persistent problems. Although deep learning strategies have greatly enhanced recognition precision, several key issues still exist concerning the lack of large-scale, diverse, and naturally emotional databases in both fields [4], [8]. The inherent subjectivity of emotional perception remains to make model assessment and standardization difficult, as emphasized by [3] and [8]. Real-time processing capacities are still limited, especially for multimodal systems that involve physiological signals [2], [7]. Potential future research directions include the creation of cross-domain representations as shown by [6]'s MTAR method, and the investigation of more sophisticated methods such as federated learning for privacy-enhancing emotion recognition [7]. The regression-based model introduced by [3] for continuous emotion prediction holds great promise for both SER and MER, and may be able to overcome limitations inherent to categorical emotion models. With advancing of the field, emphasis should be placed on developing standardized test protocols, generating more varied and ecologically valid affect datasets, and increasing the interpretability of deep learning models to enable practical deployment in applications from mental health monitoring to human-computer interaction [1], [4], [7], [8]. The incorporation of music theory principles into SER [6] and the use of probabilistic techniques [5] are especially new directions that have the potential to lead future advances in affective computing.

III. Methodology

This research plans to create an integrated machine learning system that identifies human emotion accurately from live speech inputs and suggests music tracks with corresponding emotional content. The intention is to increase user interaction through a personalized, emotion-based music recommendation experience that reflects the user's immediate emotional state. The Fig1 depicts a two-stage emotion-targeted music recommendation system in which user voice and song audio inputs are feature extracted and classified as emotions by using deep learning models. The emotion labels so predicted are used for music playing and playlist control to facilitate personalized, emotion-synchronized music sessions with real-time voice response.

A. Data Description

Two base datasets are used in this research: the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) and the MER500 music dataset. The RAVDESS dataset includes 1,440 recordings of speech uttered by 24 professional actors (12 men and 12 women) corresponding to eight emotional categories: neutral, calm, happy, sad, angry, fearful, disgust, and surprised. The dataset offers high-quality labeled speech samples used for supervised training of the speech emotion recognition module. The MER500 dataset is comprised of 500 South Indian songs, each of them annotated with one of the five emotional tags: Romantic, Party, Sad, Devotional, or Happy. The dataset can be used to train the music emotion classification module needed for emotion-based recommendation. This bar graph in Fig 2 depicts the prevalence of emotional categories in the RAVDESS speech dataset. It indicates that the 'neutral' emotion is at the highest, with nearly 290 samples, showing a greater coverage of neutral

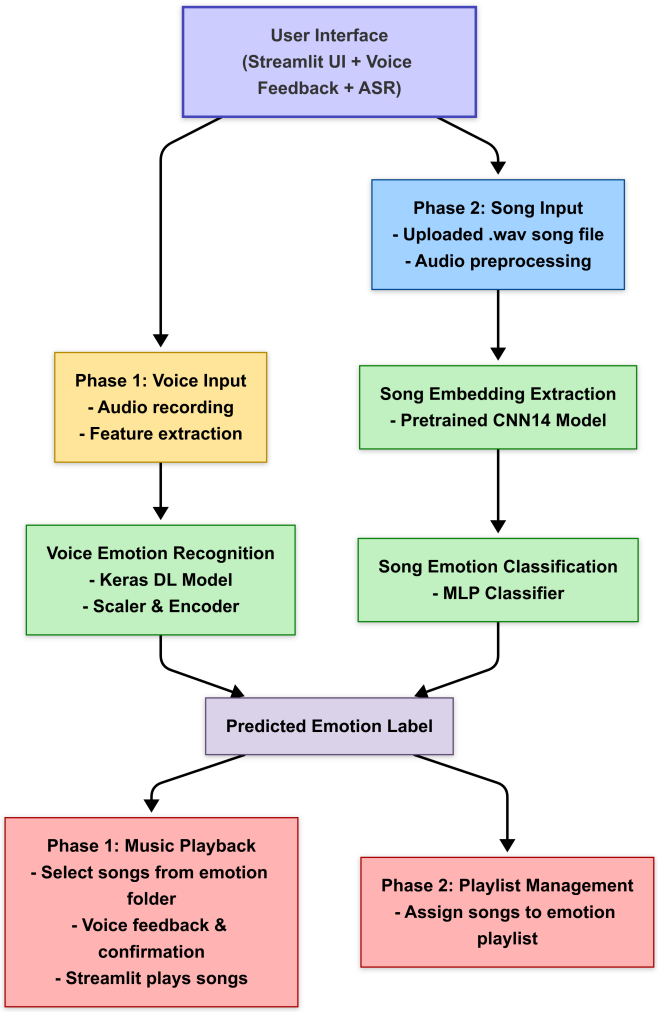

Fig. 1. Architecture diagram of Melodic Mind

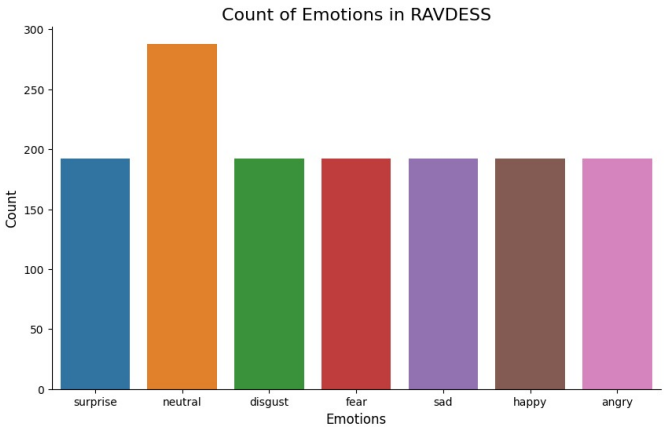

Fig. 2. Data label distribution chart of RAVDESS Dataset

speech data. The rest of the emotions—surprise, disgust, fear, sad, happy, and angry—are comparatively even, with each being approximately 190 to 195 samples. This well-balanced distribution over the majority of emotions facilitates strong training of speech emotion recognition models with ample samples for every emotional class, while a slightly greater number of neutral samples could be an indication of its natural dominance in normal speech. This bar chart in Fig 3 shows the observation counts of each emotion class in the MER500 music dataset. The distribution is extremely balanced, with every emotion class—angry, happy, neutral, romantic, and sad—being assigned around 99 to 100 samples. This very balanced dataset is useful for training machine learning models since it minimizes bias toward any specific emotion class and facilitates more robust and unbiased classification performance on every category.

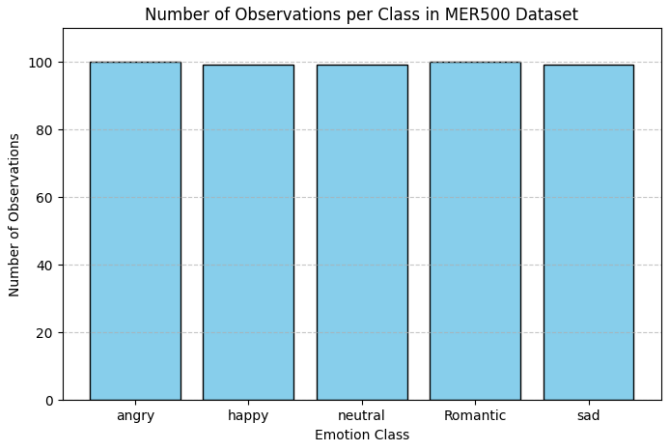

Fig. 3. Data label distribution chart of MER Dataset

B. Data Preprocessing

The speech audio files were initially trimmed to delete silence at the beginning and end with the librosa.effects.trim() function, concentrating analysis on the active speech regions. The trimmed signals were transformed into log-mel spectrograms, which retain both time and frequency features of the audio, which is optimal for deep learning feature extraction. The pre-trained CNN14 model of the PANNs framework was subsequently used to obtain informative audio embeddings from these spectrograms. For music data, a dense collection of acoustic features was calculated, such as Mel-Frequency Cepstral Coefficients (MFCCs), chroma features, spectral contrast, tempo, zero-crossing rate, spectral centroid, and roll-off. In order to treat class imbalance of the MER500 dataset—particularly for rarer classes such as Devotional. Lastly, all the feature vectors were standardized through mean-centering and variance scaling to scale feature distributions and ease model convergence.

C. Model Architecture and Training

The emotion recognition pipeline of the speech uses the pretrained CNN14 feature extractor in conjunction with a

Multi-Layer Perceptron (MLP) classifier. The CNN14 model maps log-mel spectrograms to high-level embeddings, and the MLP makes predictions on these embeddings for emotions happy, sad, angry, fear, surprise, and neutral. The MLP was trained with categorical cross-entropy loss optimized with the Adam optimizer. To achieve generalizability across speakers, Group K-Fold cross-validation was employed where data were divided by speaker identity such that all samples from the same speaker never occur in both training and test sets. Model performance was measured using metrics such as accuracy, precision, recall, and F1-score with macro-averaging.

Speech emotion recognition system based on deep learning. The audio data is preprocessed first by removing silence and extracting the Mel-frequency cepstral coefficients (MFCCs) to model significant spectral characteristics of speech. The features are normalized to provide uniform input to the model. A Convolutional Neural Network (CNN) is subsequently trained on the extracted features to provide classification of emotions from speech recordings. The table I custom architecture of the proposed CNN model. The model employs ReLU activation and dropout regularization and is trained with categorical cross-entropy loss and the Adam optimizer. For measuring model performance and preventing overfitting, K-Fold cross-validation is utilized, using accuracy, precision, recall, and F1-score as metrics for evaluation. This procedure efficiently extracts emotional features from speech and is the basis for future emotion-based applications.

Music emotion classification used an ensemble of four models: XGBoost, Random Forest, Support Vector Machine (SVM), and MLP. Each was trained in isolation on the acoustic features of the MER500 dataset. The ensemble predictions were averaged through majority voting to enhance robustness and avoid the overfitting that a single model might experience. The last system combines the output of the speech emotion recognition and music emotion classification modules. The user's speech predicted emotion serves as a query to fetch music tracks labelled with the corresponding emotional tag. Integration is done as a Flask web-based application with real-time interaction support for the user. Users input speech through microphone; the speech is processed using the CNN14 and MLP pipeline to predict the emotional state, which is employed to generate user-specific music recommendations based on the ensemble-classified MER500 dataset. The multimodal, emotionally driven recommendation framework improves user experience by issuing context-aware and emotionally consistent music selections.

TABLE I
CUSTOM CNN ARCHITECTURE

Layer (type)	**Description**
Input	Shape = (`x_train_cnn.shape[1]`, 1)
Convolutional Block 1	
Conv1D	512 filters, kernel size = 5, ReLU, padding = 'same'
BatchNormalization	-
MaxPooling1D	pool size = 5, stride = 2, padding = 'same'
Convolutional Block 2	
Conv1D	512 filters, kernel size = 5, ReLU, padding = 'same'
BatchNormalization	-
MaxPooling1D	pool size = 5, stride = 2, padding = 'same'
Dropout	rate = 0.2
Convolutional Block 3	
Conv1D	256 filters, kernel size = 5, ReLU, padding = 'same'
BatchNormalization	-
MaxPooling1D	pool size = 5, stride = 2, padding = 'same'
Convolutional Block 4	
Conv1D	256 filters, kernel size = 3, ReLU, padding = 'same'
BatchNormalization	-
MaxPooling1D	pool size = 5, stride = 2, padding = 'same'
Dropout	rate = 0.2
Convolutional Block 5	
Conv1D	128 filters, kernel size = 3, ReLU, padding = 'same'
BatchNormalization	-
MaxPooling1D	pool size = 3, stride = 2, padding = 'same'
Dropout	rate = 0.2
Flatten	-
Dense	512 units, ReLU
BatchNormalization	-
Dense	7 units, Softmax (output layer for 7 emotion classes)

IV. RESULTS AND DISCUSSION

This scatter plot Fig 4 displays a two-dimensional t-SNE representation of audio embeddings obtained with the pretrained CNN14 model from the PANNs framework. Various colors symbolize different emotion classes (angry, happy, neutral, romantic, sad), demonstrating the clustering and separability of audio features in terms of emotional content. The plot suggests that the model is able to catch significant emotional differences in the latent space, where there is some overlap among classes. This plot Fig 5 shows Receiver Operating

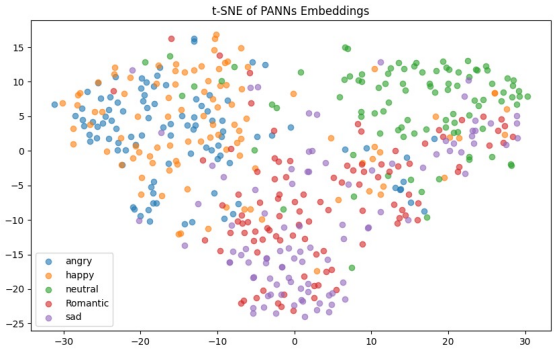

Fig. 4. t-SNE visualization of pretrained CNN14 embeddings for music tracks colored by emotion class

Characteristic (ROC) curves of the same music emotion classification model, with the true positive rate plotted against the false positive rate at various thresholds. The large AUC values (from 0.87 to 0.97) across classes show good discrimination ability of the ensemble classifiers, and the best performance is from the 'neutral' emotion. The graph Fig 6 illustrates precision-recall curves for every emotion class in the task

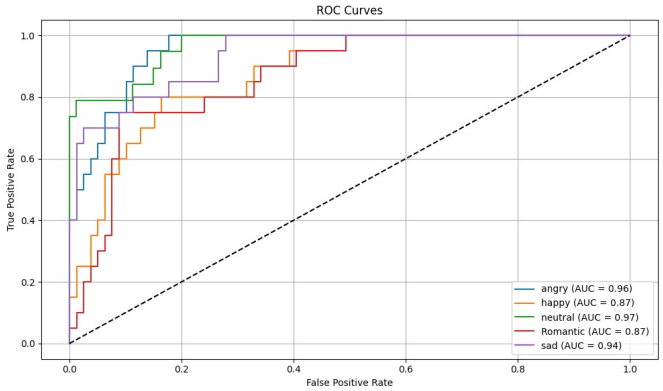

Fig. 5. ROC curves for music emotion classification showing classifier sensitivity and specificity

of music emotion classification. Every curve represents the trade-off between precision and recall at different classification thresholds. The Area Under Curve (AUC) measures classifier performance, as 'neutral' and 'angry' classes reach the highest scores (0.91 and 0.84 respectively), whereas 'romantic' has the worst AUC (0.57), implying difficulty with classification of that category.

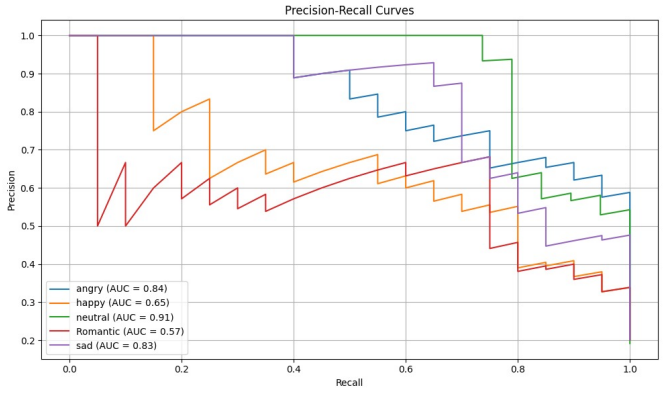

Fig. 6. Precision-Recall curves for music emotion classification across different emotion classes

The plot in Fig 7 shows training and validation loss and accuracy curves for 30 epochs of a deep learning model. Training loss falls continuously to close to zero, whereas validation loss falls initially but levels off, showing some overfitting. Training accuracy approaches 100%, while validation accuracy plateaus at around 85%, indicating good but not perfect generalization to novel data.

The table II illustrates the performance of two models—Random Forest and CNN—on the task of speech emotion classification. The CNN model performs much better than Random Forest, with 85% precision, 84% recall, 84% F1 score, and 84% accuracy. By comparison, the Random Forest model has much lower values, precision at 43%, recall at 41%, F1 score at 37%, and accuracy at 44%, suggesting CNN is better suited for speech emotion recognition in this research.

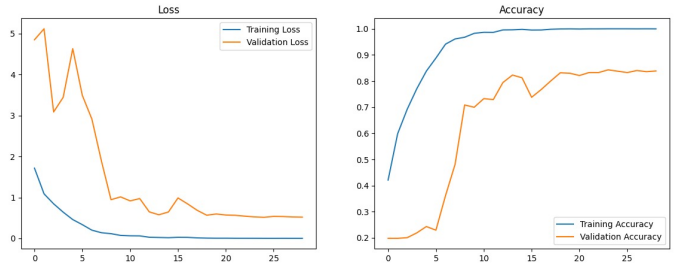

Fig. 7. Training and validation loss and accuracy curves of person emotion recognition model

TABLE II
MODEL PERFORMANCE OF SPEECH CLASSIFICATION

Models	Precision	Recall	F1 score	Accuracy
CNN	85%	84%	84%	84%
Random forest	43%	41%	37%	44%

The table III is the classification performance of three methods—MLP, ensemble method, and Random Forest—on the music emotion classification task. MLP has the best performance with 83% precision and recall, 83% F1 score, and 84% accuracy. The ensemble method demonstrates moderate performance with approximately 61% on precision, recall, F1 score, and accuracy. The Random Forest model is behind with results in mid-50%. This indicates that MLP achieves better music emotion classification than the other tested schemes.

TABLE III
MODEL PERFORMANCE OF MUSIC CLASSIFICATION

Class	Precision	Recall	F1 Score	Accuracy
MLP	83%	83%	83%	84%
Ensemble	61%	62%	61%	62%
Random forest	55%	55%	55%	56%

V. CONCLUSION

Melodic Mind has effectively created a multimodal emotion-sensitive music recommendation framework by combining speech emotion recognition and music emotion classification across three emotional categories: valence, arousal, and dominance. It shows that using pre-trained deep audio feature extractors (CNN14 from PANNs) together with a Multi-Layer Perceptron for speech emotion recognition achieves strong classification results across multiple emotional categories. Likewise, ensemble learning approaches applied to the MER500 music collection across three emotional categories achieved accurate music emotion classification, facilitating effective song to user emotion matching from speech.

Its design, involving strict data preprocessing, and validation methods such as Group K-Fold and stratified cross-validation, provides for generalizability and stability in practice usage. Performance metrics of precision, recall, F1-score, and ROC/AUC curves validate good model performance, although some emotion classes—most notably more subjective ones such as 'romantic'—are more difficult to classify. The real-time recommendation pipeline based on Flask illustrates

the applied relevance of this paradigm in context-aware and personalized music recommendation with prospects for entertainment, mental well-being, and therapy applications.

Nonetheless, there are limitations such as mid-range variance in performance across emotion classes and slight overfitting apparent from training versus validation curves, implying that there could be potential improvements by increasing datasets, more advanced model architectures, or using multimodal signals other than audio, i.e., facial expressions or physiological signals. Real-time adaptive learning, cross-cultural emotion modeling, and integration with larger affective computing frameworks could be potential directions of investigation for future studies. In addition to that, examining feedback mechanisms of users and longitudinal experiments on the influence of emotion-based music recommendation on mood control would be informative.

In summary, this research presents a novel, end-to-end system connecting speech emotion analysis and music emotion classification, pushing the frontiers of affective computing and personal recommendation systems

REFERENCES

[1] Ahlam Hashem, Muhammad Arif, Manal Alghamdi, "Speech emotion recognition approaches: A systematic review", Speech Communication, Volume 154, 2023, 102974, ISSN 0167-6393, https://doi.org/10.1016/j.specom.2023.102974.

[2] Xingye Hao, Honghe Li, Yonggang Wen, "Real-time music emotion recognition based on multimodal fusion", Alexandria Engineering Journal, Volume 116, 2025, Pages 586-600, ISSN 1110-0168, https://doi.org/10.1016/j.aej.2024.12.060.

[3] Y. -H. Yang, Y. -C. Lin, Y. -F. Su and H. H. Chen, "A Regression Approach to Music Emotion Recognition", in IEEE Transactions on Audio, Speech, and Language Processing, vol. 16, no. 2, pp. 448-457, Feb. 2008, doi: 10.1109/TASL.2007.911513.

[4] X. Jiang, Y. Zhang, G. Lin and L. Yu, "Music Emotion Recognition Based on Deep Learning: A Review", in IEEE Access, vol. 12, pp. 157716-157745, 2024, doi: 10.1109/ACCESS.2024.3484470.

[5] K. Markov and T. Matsui, "Music Genre and Emotion Recognition Using Gaussian Processes", in IEEE Access, vol. 2, pp. 688-697, 2014, doi: 10.1109/ACCESS.2014.2333095.

[6] X. Li et al., "Music Theory-Inspired Acoustic Representation for Speech Emotion Recognition", in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, pp. 2534-2547, 2023, doi: 10.1109/TASLP.2023.3289312.

[7] Kheddar, Hamza & Hemis, Mustapha & Himeur, Yassine. (2024). "Automatic speech recognition using advanced deep learning approaches: A survey. Information Fusion". 109. 102422. 10.1016/j.inffus.2024.102422.

[8] Han, D., Kong, Y., Han, J. et al. "A survey of music emotion recognition". Front. Comput. Sci. 16, 166335 (2022). https://doi.org/10.1007/s11704-021-0569-4

[9] Bhangale, Kishor & Kothandaraman, Mohanaprasad. (2023). "Speech Emotion Recognition Based on Multiple Acoustic Features and Deep Convolutional Neural Network". Electronics. 12. 839. 10.3390/electronics12040839.

[10] Samaneh Madanian, Talen Chen, Olayinka Adeleye, John Michael Templeton, Christian Poellabauer, Dave Parry, Sandra L. Schneider, "Speech emotion recognition using machine learning — A systematic review", Intelligent Systems with Applications, Volume 20, 2023, 200266, ISSN 2667-3053, https://doi.org/10.1016/j.iswa.2023.200266.

[11] M. S. Guthula and M. Bordoloi, "Music Recommendation System Using Facial Detection Based Emotion Analysis," 2024 International Conference on Emerging Techniques in Computational Intelligence (ICETCI), Hyderabad, India, 2024, pp. 296-301, doi: 10.1109/ICETCI62771.2024.10704201.

[12] Z. Anna, "Research on Music Emotion Recognition and Intelligent Classification System Based on Deep Learning," 2024 IEEE 2nd International Conference on Sensors, Electronics and Computer Engineering (ICSECE), Jinzhou, China, 2024, pp. 1061-1065, doi: 10.1109/ICSECE61636.2024.10729602.

[13] M. Guo, "Intelligent Classification of Music Emotions Based on Multi-feature Fusion Algorithm of Feedforward Neural Network," 2024 International Conference on Electrical Drives, Power Electronics Engineering (EDPEE), Athens, Greece, 2024, pp. 766-769, doi: 10.1109/EDPEE61724.2024.00147.

[14] P. Du, X. Li and Y. Gao, "Dynamic Music Emotion Recognition Based on CNN-BiLSTM," 2020 IEEE 5th Information Technology and Mechatronics Engineering Conference (ITOEC), Chongqing, China, 2020, pp. 1372-1376, doi: 10.1109/ITOEC49072.2020.9141729.

[15] V. R. Revathy and A. S. Pillai, "Multi-class Classification of Song Emotions Using Machine Learning," 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 2022, pp. 2317-2322, doi: 10.1109/ICACITE53722.2022.9823535.

[16] Amrita Nair, Smriti Pillai, Ganga S. Nair, and T. Anjali, "Emotion Based Music Playlist Recommendation System Using Interactive Chatbot," 2021 6th International Conference on Communication and Electronics Systems (ICCES), pp. 1767-1772, IEEE, 2021.

[17] K. Seshaayani, Srinithya SL, Pallavi, Visalatchi, Neelima, "Emotion Recognition Based Music Player," Fifth International Conference on Electrical, Computer and Communication Technologies (ICECCT), 2023, pp. 1-5.

[18] K. R. Nambiar and S. Palaniswamy, "Speech Emotion Based Music Recommendation," 2022 3rd International Conference for Emerging Technology (INCET), 2022.

[19] C. Selvi and E. Sivasankar, "An Efficient Context-Aware Music Recommendation Based on Emotion and Time Context," Lecture Notes on Data Engineering and Communications Technologies, Indore, 2018.

[20] Trisha, Kariveda, Padigela Srinithya Reddy, Nichenametla Hima Sree, Tripty Singh, and Mansi Sharma, "Deep Learning-Enhanced Emotion-Based Music System with Age and Language Personalization," 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), pp. 1-7, IEEE, 2024.