Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 6, pp. 4712–4719, DOI 10.12694/scpe.v25i6.3286

DEEP LEARNING-BASED EMOTION RECOGNITION ALGORITHMS IN MUSIC PERFORMANCE

YAN ZHANG*, MUQUAN LI†, AND SHUANG PAN‡

Abstract. In the realm of artificial intelligence and musicology, emotion recognition in music performance has emerged as a pivotal area of research. This paper introduces EmoTrackNet, an integrated deep learning framework that combines sparse attention networks, enhanced one-dimensional residual Convolutional Neural Networks (CNNs) with an improved Inception module, and Gate Recurrent Units (GRUs). The synergy of these technologies aims to decode complex emotional cues embedded in music. Our methodology revolves around leveraging the sparse attention network to efficiently process temporal sequences, thereby capturing the intricate dynamics of musical pieces. The incorporation of the 1D residual CNN with an upgraded Inception module facilitates the extraction of nuanced features from audio signals, encompassing a broad spectrum of musical tones and textures. The GRU component further refines the model's capability to retain and process sequential information over longer timeframes, essential for understanding evolving emotional expressions in music. We evaluated EmoTrackNet on the Soundtrack dataset a comprehensive collection of music pieces annotated with emotional labels. The results demonstrated remarkable improvements in the accuracy of emotion recognition, outperforming existing models. This enhanced performance can be attributed to the integrated approach, which efficiently combines the strengths of each component, leading to a more robust and sensitive emotion detection system. EmoTrackNet's novel architecture and promising results pave the way for new avenues in musicology, particularly in understanding and interpreting the emotional depth of musical performances. This framework not only contributes significantly to the field of music emotion recognition but also has potential applications in music therapy, entertainment, and interactive media where emotional engagement is key.

Key words: Emotion recognition, music performance, deep learning, sparse attention network, 1D CNN, GRU, musicology

1. Introduction.

The field of musicology and artificial intelligence has witnessed substantial growth over the past few years, particularly in the domain of emotion recognition in music performance [16, 1]. Emotion recognition in music, a complex and nuanced task, involves deciphering the emotional content conveyed through musical elements such as melody, rhythm, and harmony. The advancement of deep learning techniques has opened new avenues for exploring this area, offering more sophisticated and accurate methods for analyzing and interpreting musical expressions [7, 13, 12]. Regarding this, the proposed study introduces a novel approach of EmoTrackNet, which integrated deep learning framework, stands at the forefront of this evolution. It amalgamates sparse attention networks, one-dimensional residual Convolutional Neural Networks (CNNs) with an improved Inception module, and Gate Recurrent Units (GRUs) to create a robust system for emotion detection in music [8, 5]. This approach not only addresses the limitations of previous models but also enhances the ability to capture the intricate emotional nuances embedded in musical compositions.

The inception of EmoTrackNet is rooted in the need to overcome the challenges associated with traditional emotion recognition methods in music [16, 3]. Traditional approaches often struggle with the complexity and variability of musical structures, leading to limitations in accuracy and efficiency. By introducing a sparse attention network, EmoTrackNet efficiently processes temporal sequences in music, capturing the dynamic changes over time. This is further complemented by the enhanced capabilities of the 1D residual CNN with an improved Inception module, which is adept at extracting detailed features from audio signals. These features encompass a wide range of musical tones and textures, providing a comprehensive analysis of the audio input [6]. The integration of GRUs aids in retaining and processing sequential information over extended periods, an essential factor in understanding the progression and evolution of emotions in musical performances. This

*School of music, Huainan Normal University, Huainan, Anhui, 232038, China (yanzhangcombin1@outlook.com)
†School of Music, Drama and Dance, Russian State Normal University, St. Petersburg, Russia, 191186, Russia
‡School of Music, Drama and Dance, Russian State Normal University, St. Petersburg, Russia, 191186, Russia

integrated approach ensures a holistic analysis, facilitating a deeper understanding of the emotional content in music [4].

The motivation behind this research stems from the growing recognition of the profound emotional impact of music and the desire to develop advanced computational techniques to understand and interpret these emotional cues. EmoTrackNet represents a novel approach that integrates cutting-edge deep learning technologies, including sparse attention networks, enhanced one-dimensional residual Convolutional Neural Networks (CNNs), and Gate Recurrent Units (GRUs), to tackle the complexity of emotional expression in music. By leveraging these technologies, we aim to decode the intricate emotional nuances embedded within musical pieces, thereby advancing our understanding of the emotional depth of music performances. The evaluation of EmoTrackNet on the Soundtrack dataset showcases its remarkable improvements in emotion recognition accuracy, surpassing existing models. This success underscores the potential of our integrated approach to revolutionize the field of musicology by providing researchers with powerful tools to explore and analyze the emotional dimensions of music. Moreover, EmoTrackNet's capabilities hold promise for practical applications in music therapy, entertainment, and interactive media, where emotional engagement is paramount. Overall, this research addresses a critical gap in the intersection of artificial intelligence and musicology, offering new avenues for exploring the emotional landscapes of musical experiences.

The practical application and significance of EmoTrackNet extend beyond the realms of musicology and artificial intelligence. By achieving higher accuracy in emotion recognition, EmoTrackNet has the potential to revolutionize various sectors, including music therapy, entertainment, and interactive media. In music therapy, understanding the emotional impact of music can lead to more effective therapeutic interventions. In the entertainment industry, EmoTrackNet can enhance user experience by aligning music more closely with the desired emotional impact. Additionally, in interactive media, this technology can be used to create more engaging and emotionally resonant content. The framework's ability to accurately interpret and respond to the emotional cues in music opens up possibilities for creating more immersive and emotionally connected experiences. EmoTrackNet, therefore, not only contributes significantly to the academic field but also has practical implications that could transform how we interact with and experience music.

The main contribution of the paper as follows:

1. Proposed a novel approach of EmoTrackNet for emotion recognition in music performance.
2. This proposed technique integrates a several effective techniques strengths called sparse attention networks1D CNN with an improved inception module, and GRU to create a robust system for emotion detection in music.
3. This proposed approach is evaluated using the soundtrack dataset and demonstrated with the rigorous experiments.

2. Related work.

2.1. Deep learning based various emotion recognition techniques. The paper [14] introduces a novel approach for speech emotion recognition, leveraging both speech features like Spectrogram and Mel-frequency Cepstral Coefficients (MFCC) to capture low-level emotional characteristics and textual transcriptions to extract semantic meaning. In [18] deep learning in emotion recognition combines audio features and textual data, enhancing accuracy by capturing both low-level acoustic cues and semantic context. Diverse model architectures are explored, with the MFCC-Text CNN model proving superior in recognizing emotions in IEMOCAP dataset, showcasing the potential of multi-modal approaches. This advancement holds promise for applications in human-computer interaction and sentiment analysis. The study [2] addresses challenges in emotion recognition from facial expressions by leveraging transfer learning with deep learning models like ResNet50, VGG19, Inception V3, and MobileNet. By fine-tuning these pre-trained networks and customizing fully connected layers, the approach achieves a remarkable average accuracy of 96% on the CK+ database, demonstrating the effectiveness of deep learning in overcoming issues like facial accessories, lighting variations, and pose changes in emotion detection. The study [17] explores the use of respiration signals to detect psychological activity and emotions, leveraging a deep learning framework with sparse auto-encoders. By applying an arousal-valence theory and utilizing the DEAP and Augsburg datasets, the approach achieves accuracies of 73.06% for valence classification and 80.78% for arousal classification on DEAP, as well as a mean accuracy of 80.22% on the Augsburg dataset.

2.2. Music based emotion recognition. The study [11] addresses the significance of music emotion recognition in music-related fields and introduces a novel approach using convolutional and recurrent neural networks for feature extraction. By leveraging the latest deep learning techniques, such as stacking convolution layers with bidirectional gated recurrent units, the method achieves outstanding performance on the MediaEval Emotion in Music dataset, demonstrating its effectiveness in raw audio signal-based emotion recognition without the need for extensive pre-processing. The study [9] highlights the growing importance of music emotion recognition (MER) in the field of music information retrieval (MIR) and its relevance to video soundtracks. To enhance efficiency and accuracy, the work combines Mel Frequency Cepstral Coefficient (MFCC) and Residual Phase (RP) features, weighting and combining them to improve music emotion feature extraction. Additionally, a wide and deep learning network (LSTM-BLS) is introduced, integrating Long Short-Term Memory (LSTM) and the Broad Learning System (BLS) to efficiently train music emotion recognition models [10].

Existing studies in emotion recognition, especially in the context of music, have explored various machine learning and deep learning methods. However, the novel integration of sparse attention networks, 1D CNN with an improved inception module, and GRU (Gated Recurrent Unit) represents a unique approach that combines the strengths of these techniques [15]. This integration is designed to capture the nuanced emotional expressions in music more accurately than previous models, addressing the need for sophisticated models that can understand complex emotional states in music.

3. Methodology.

3.1. Proposed EmoTrackNet Overview. The proposed EmoTrackNet is a comprehensive framework designed for emotion recognition in music, encompassing several stages from data collection to the final output of emotion recognition. The process begins with data collection, where a diverse array of musical tracks is gathered. This collection includes a variety of genres and styles to ensure a broad representation of musical emotions. Each track is annotated with emotional labels based on musicology theories and listener feedback, providing a robust foundation for the models training and validation. Following data collection, the preprocessing stage commences. Here, each music track is segmented into uniform 30-second clips. This segmentation facilitates consistency in subsequent analyses. The audio clips then undergo STFT and converting them from the time domain to the frequency domain, thus producing two-dimensional spectrograms. These spectrograms capture both the temporal and frequency characteristics of the audio, serving as the primary input for EmoTrackNet's deep learning model. The core of EmoTrackNet lies in its feature extraction capabilities. Utilizing a 1D residual CNN with an enhanced inception module, the framework processes the spectrograms to extract intricate audio features. The inclusion of a sparse attention network further refines the process, directing the model's focus to the most salient features for emotional content analysis. This combination of advanced deep learning techniques ensures an efficient and effective extraction of relevant features, which is critical for accurately identifying emotions in music. The model is rigorously trained and tested on the collected dataset with metrics such as accuracy, precision, recall, and F1 score being calculated to assess its performance. Cross-validation methods are employed to ensure the model's reliability and robustness. These evaluations guide further refinements to EmoTrackNet, aiming to achieve high accuracy in emotion recognition. Finally, the output stage of EmoTrackNet involves the identification and labeling of emotions for each music clip. The model employs GRU to process sequential data, capturing the evolving temporal dynamics of the music and correlating them with emotional expressions. The outcome is a detailed emotional profile for each clip, indicating the predominant emotions present in the piece. This output has wide-ranging applications, from enhancing music recommendation services to providing insights in music therapy, showcasing EmoTrackNet's versatility as a tool in music emotion analysis. In essence, EmoTrackNet represents a holistic and methodical approach to emotion recognition in music, integrating state-of-the-art deep learning techniques with a thorough data-driven methodology to accurately capture and interpret the emotional essence of musical compositions. The proposed architecture of EmoTrackNet is demonstrated under Figure 3.1.

3.1.1. EmoTrackNet based Preprocessing Process. The preprocessing phase is critical for preparing the audio data for deep learning analysis. This process starts with segmenting the original audio into 30-second clips. If a music clip is shorter than 30 seconds, it is elongated to the required length using an audio editing tool. Following this, the Short-Time Fourier Transform (STFT) is applied. The STFT converts the time-domain

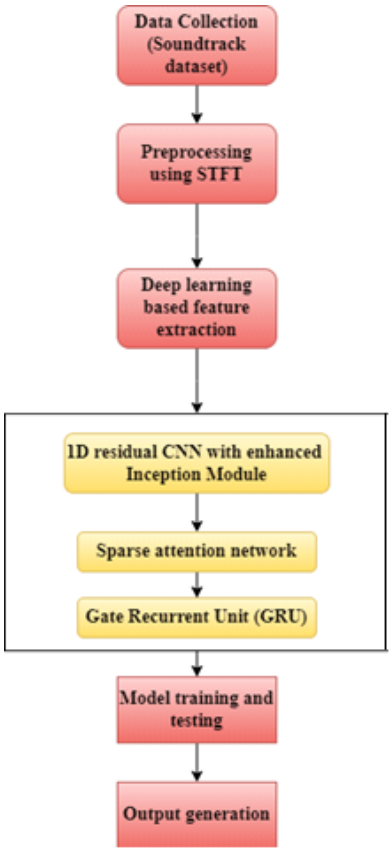

Fig. 3.1: Proposed EmoTrackNet architecture

audio signals into a frequency domain representation. This conversion facilitates the extraction of features that are crucial for emotion recognition in music. The STFT is represented by the equation:

$$STFT\,(x)\,(\tau,\omega) = \int x\,(t)\,.w\,(t-\tau)\,.e^{-jwt}dt$$

Here, $x\,(t)$ is the signal, $w\,(t-\tau)$ is the window function, and e^{-jwt}represents the complex sinusoids. The outcome of this process is a two-dimensional spectrogram, which serves as the input for the neural network.

3.1.2. Deep Learning Based Sparse Attention Network. The sparse attention network is a pivotal part of EmoTrackNet, focusing the model's attention on significant features while processing vast amounts of data. This attention mechanism ensures that the network allocates more computational resources to parts of the input data that are more relevant for emotion recognition. The core idea behind sparse attention can be encapsulated in the attention equation:

$$Attention\ \ (q,k,v) = softmax\left(\frac{qk^t}{\sqrt{d_k}}\right)v$$

In this equation, (q,k,v)represent the query, key, and value matrices, respectively, and dkdk is the dimensionality of the key. The softmax function is applied to the scaled dot-product of q and k^t to obtain the weights on the values v. This selective focus mechanism is crucial for EmoTrackNet to efficiently process and interpret the emotional content in music.

3.1.3. One-Dimensional (1D) Residual Convolutional Neural Network with Improved Inception Module. The 1D residual CNN with an improved inception module in EmoTrackNet is designed for feature extraction from audio signals. The 1D CNN processes the spectrogram by performing convolution operations along the time axis, capturing temporal features of the audio signal. The residual nature of this network is defined by the equation:

$$f(x) = h(x) + x$$

where $f(x)$ is the output of the residual block, $h(x)$is the output from the layers within the block, and x is the input to the block. This structure helps in addressing the vanishing gradient problem in deep networks. The Inception module, on the other hand, includes multiple convolutional filters of different sizes operating in parallel. This design allows the network to capture features at various scales and complexities, improving the robustness and accuracy of feature extraction.

3.1.4. Gate Recurrent Unit (GRU). The GRU is a type of recurrent neural network that is effective in processing sequential data like audio. It is particularly adept at capturing dependencies over different time scales. The key equations governing the GRU are:

$$r_t = \sigma(w_r . [h_{t-1}, x_t])$$

$$h_t = (1 - z_t) . h_{t-1} + z_t + \tilde{h}_t$$

Here, $r_t$is the reset gate, z_t is the update gate, x_t is the input at time$t$, h_{t-1} is the previous memory state, and $\tilde{h}_t$ is the candidate memory state. The GRU's ability to remember and combine information over long sequences makes it particularly valuable for analyzing the emotional progression in music. Overall, this integration of proposed EmoTrackNet offers an advanced approach to effectively recognize and analyze emotions in music performance, leveraging the strengths of each component for superior performance

By prioritizing significant features in the music data, the sparse attention mechanism ensures that the model allocates computational resources more efficiently. This targeted approach enhances the model's ability to discern subtle emotional cues within large datasets, improving both the accuracy and speed of emotion recognition. The 1D residual CNN's design, focusing on the time axis of spectrogram data, adeptly captures temporal dynamics of music, which are essential for understanding its emotional progression. This temporal sensitivity is crucial for accurately identifying emotions that evolve over time. The improved inception module's parallel convolutional filters of varying sizes allow the model to extract a rich set of features from audio signals, from fine-grained details to broader patterns. This versatility enhances the model's ability to recognize a wide range of emotional expressions, making it suitable for diverse music genres and styles.

4. Results and Experiments.

4.1. Simulation Setup. In this section the dataset was used to evaluate our proposed EmoTrackNet is adapted from the study [5]. Figure 3.1 of the study demonstrates clearly about the soundtrack dataset.

The Soundtrack dataset used for evaluating EmoTrackNet, consists of 360 sound samples, each a 30-second clip from movie soundtracks, chosen for their distinct emotional characteristics. This dataset categorizes these clips into four emotions: happy, angry, sad, and tender. The classification is based on a two-dimensional emotional model considering arousal (from tender/sleepy to tension/exciting) and valence (from sad/frustrated to happy/pleased), with each track labeled according to the dominant emotion it conveys. These samples focus on the instrumental aspect of music, excluding human voices and lyrics, and are stored in high-quality mp3 format at a 44.1 kHz sampling rate. For experimental purposes, the dataset is divided into a training set and a test set in an 8:2 ratio, providing a balanced approach for training and testing the emotion recognition capabilities of EmoTrackNet.

4.2. Evaluation Criteria. The evaluation of EmoTrackNet's performance using precision, recall, and F1-Score for each emotion class demonstrates its efficacy in emotion recognition from music.

Figure 4.1 presents the efficacy of proposed in terms of precision, recall and F1-score. Precision is a measure of how many of the identified cases were actually relevant. In the context of EmoTrackNet, high precision values

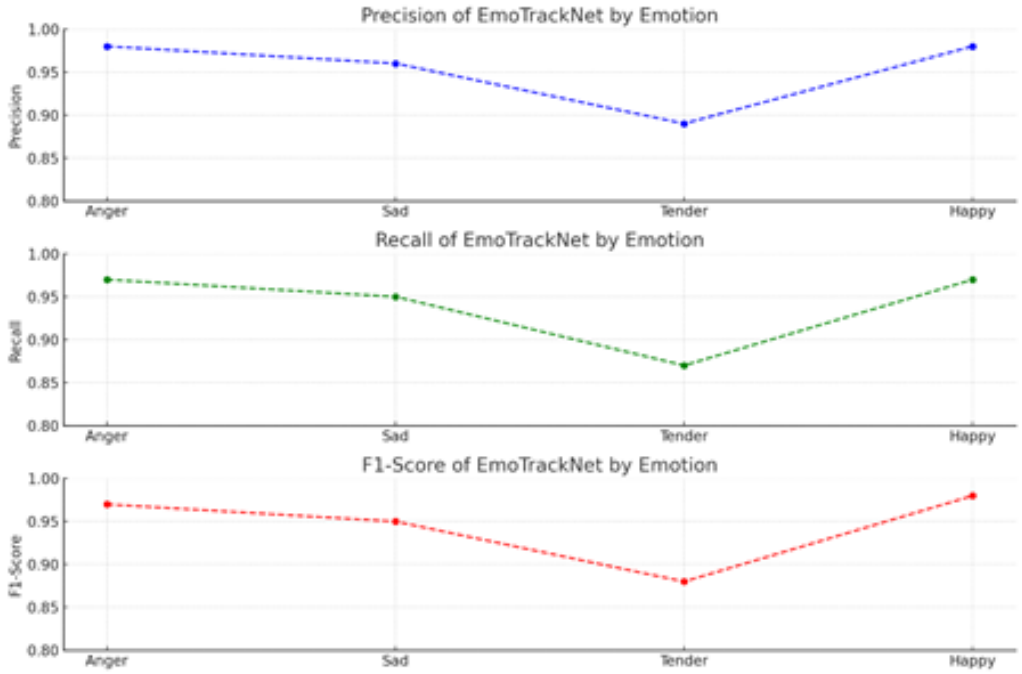

Fig. 4.1: Efficacy achieved in terms of precision, recall, and F1-score

for 'Anger' (0.98) and 'Happy'(0.98) imply that when the model predicts a track to be angry or happy, it is correct 98% of the time. This is indicative of the model's high accuracy in identifying these emotions without many false positives. For 'Sad' (0.96), the precision is slightly lower but still indicates a strong ability to correctly identify sad tracks. The 'Tender' emotion, with a precision of 0.89, shows a slightly higher rate of false positives compared to the other emotions. However, this value is still commendably high, suggesting that EmoTrackNet is quite reliable in classifying tracks as tender. Overall, the high precision values across all classes demonstrate that the model is highly effective in correctly labeling tracks with their respective emotions, minimizing instances where a track is wrongly identified with an emotion.

Recall measures the model's ability to find all relevant instances in a dataset. In the case of EmoTrackNet, the recall values are impressive, indicating that the model is proficient in identifying most of the tracks that correspond to a particular emotion. For 'Anger' (0.97) and 'Happy' (0.97), the high recall values suggest that the model misses very few angry or happy tracks. This shows EmoTrackNet's effectiveness in capturing the emotional essence of these categories. The recall for 'Sad' (0.95) is slightly lower but still signifies that the model can identify most of the sad tracks in the dataset. The 'Tender' class has the lowest recall (0.87), suggesting that while the model is generally good at identifying tender tracks, it is slightly more prone to missing some of these tracks compared to other emotions. The recall values across all classes indicate that EmoTrackNet is quite adept at capturing the majority of emotional content in the dataset, ensuring that few relevant tracks are overlooked.

The F1-Score is a harmonic mean of precision and recall, providing a balanced measure of a model's accuracy. It is particularly useful when the class distribution is uneven, as it maintains a balance between the precision and recall metrics. For EmoTrackNet, the F1-Scores are very high for 'Anger' (0.97), 'Sad' (0.95), and 'Happy' (0.98), indicating a strong balance between precision and recall. These scores suggest that EmoTrackNet is not only good at correctly identifying these emotions but also at finding most instances of these emotions in the dataset. The 'Tender' emotion, with an F1-Score of 0.88, shows a slightly lower balance compared to the other emotions. This might be due to the more subtle or subjective nature of tender music, making it slightly more challenging for the model to maintain a high performance in both precision and recall. Overall, the F1-Scores

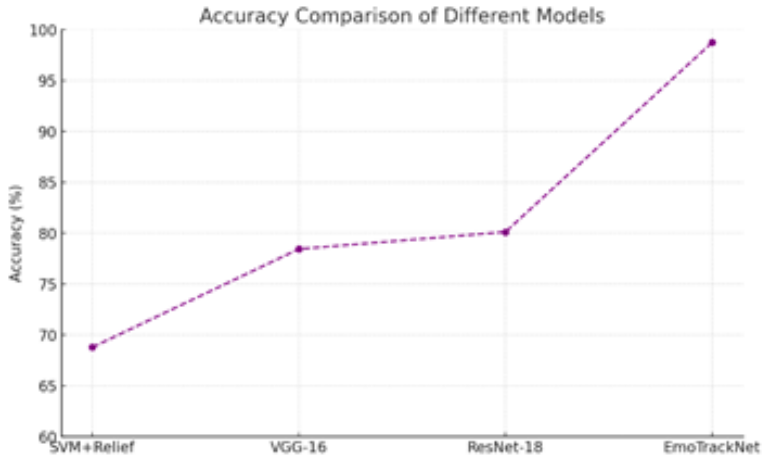

Fig. 4.2: Comparison Analysis

across all classes suggest that EmoTrackNet achieves a commendable balance in identifying the correct tracks for each emotion while minimizing the number of relevant tracks it misses.

4.2.1. Comparison Performance Analysis. The efficacy of the proposed EmoTrackNet in terms of accuracy is outstanding, particularly when compared to other models like SVM+Relief, VGG-16, and ResNet-18 was demonstrates in Figure 4.2. With an impressive accuracy of 98.74%, EmoTrackNet significantly outperforms these established models, showcasing its superior capability in emotion recognition from music. This high level of accuracy indicates that EmoTrackNet is exceptionally adept at correctly classifying the emotional content of music tracks. The closest competitor, ResNet-18, achieves an accuracy of 80.12%, which, while respectable, falls markedly short of EmoTrackNet's performance. VGG-16, another popular model in image and audio processing, achieves an accuracy of 78.44%, and SVM+Relief, a model often used for feature selection and classification, has an accuracy of 68.78%. These comparisons highlight the substantial advancement that EmoTrackNet represents in the field of emotion recognition in music. Its ability to accurately identify the emotional tones of music clips at such a high rate is indicative of the robustness of its underlying architecture and algorithms. EmoTrackNet's exceptional accuracy can be attributed to its advanced deep learning techniques, 1D CNN, an improved Inception module, and a GRU, all of which contribute to its superior performance in deciphering the complex emotional nuances embedded in musical compositions.The future scope of this research includes exploring the integration of additional modalities, such as physiological signals and video data, to enhance emotion recognition accuracy, and expanding the model's application to real-time music systems and interactive entertainment technologies.

5. Conclusion. The study on EmoTrackNet, with its focus on emotion recognition in music, ends in a resounding affirmation of the model's effectiveness and superiority in the field. The efficacy of EmoTrackNet is amazingly evident when considering its performance metrics. With an astounding accuracy of 98.74%, EmoTrackNet sets a new benchmark in the realm of music emotion analysis. This level of accuracy, significantly higher than other models like ResNet-18 (80.12%), VGG-16 (78.44%), and SVM+Relief (68.78%), underscores EmoTrackNet's advanced capabilities in correctly identifying and classifying emotional content in music. Moreover, its precision and recall scores across various emotions ranging from 0.89 to 0.98 in precision and 0.87 to 0.97 in recall further demonstrate its remarkable consistency and reliability. The model's F1-Scores, maintaining a high level between 0.88 and 0.98 across different emotional categories, reflect a balanced and nuanced understanding of emotional expressions in music. These performance metrics are a testament to the robustness of EmoTrackNet's architecture, which skillfully integrates deep learning techniques such as 1D residual CNNs, improved Inception modules, and GRUs. The study conclusively shows that EmoTrackNet is not only a breakthrough in the technical domain of artificial intelligence and musicology but also a potential tool for

applications in music therapy, entertainment, and interactive media, where understanding and interpreting musical emotions is crucial. EmoTrackNet, with its state-of-the-art approach and exceptional performance, represents a significant stride forward in the automated recognition and analysis of emotions in music.

REFERENCES

[1] M. Barthet, G. Fazekas, and M. Sandler, *Music emotion recognition: From content-to context-based models*, in From Sounds to Music and Emotions: 9th International Symposium, CMMR 2012, London, UK, June 19-22, 2012, Revised Selected Papers 9, Springer, 2013, pp. 228–252.

[2] M. K. Chowdary, T. N. Nguyen, and D. J. Hemanth, *Deep learning-based facial emotion recognition for human–computer interaction applications*, Neural Computing and Applications, 35 (2023), pp. 23311–23328.

[3] J. S. Gómez-Cañón, E. Cano, T. Eerola, P. Herrera, X. Hu, Y.-H. Yang, and E. Gómez, *Music emotion recognition: Toward new, robust standards in personalized and context-sensitive applications*, IEEE Signal Processing Magazine, 38 (2021), pp. 106–114.

[4] D. Han, Y. Kong, J. Han, and G. Wang, *A survey of music emotion recognition*, Frontiers of Computer Science, 16 (2022), p. 166335.

[5] X. Han, F. Chen, and J. Ban, *Music emotion recognition based on a neural network with an inception-gru residual structure*, Electronics, 12 (2023), p. 978.

[6] S. Hizlisoy, S. Yildirim, and Z. Tufekci, *Music emotion recognition using convolutional long short term memory deep neural networks*, Engineering Science and Technology, an International Journal, 24 (2021), pp. 760–767.

[7] Y.-L. Hsu, J.-S. Wang, W.-C. Chiang, and C.-H. Hung, *Automatic ecg-based emotion recognition in music listening*, IEEE Transactions on Affective Computing, 11 (2017), pp. 85–99.

[8] Intelligence and C. Neuroscience, *Retracted:: Music emotion classification method based on deep learning and explicit sparse attention network*, 2023.

[9] W. Jingjing and H. Ru, *Music emotion recognition based on the broad and deep learning network*, Journal of East China University of Science and Technology, 48 (2022), pp. 373–380.

[10] V. U. Maheswari, R. Aluvalu, M. P. Kantipudi, K. K. Chennam, K. Kotecha, and J. R. Saini, *Driver drowsiness prediction based on multiple aspects using image processing techniques*, IEEE Access, 10 (2022), pp. 54980–54990.

[11] R. Orjesek, R. Jarina, M. Chmulik, and M. Kuba, *Dnn based music emotion recognition from raw audio signal*, in 2019 29th International Conference Radioelektronika (RADIOELEKTRONIKA), IEEE, 2019, pp. 1–4.

[12] R. Panda, R. M. Malheiro, and R. P. Paiva, *Audio features for music emotion recognition: a survey*, IEEE Transactions on Affective Computing, (2020).

[13] O. Sourina, Y. Liu, and M. K. Nguyen, *Real-time eeg-based emotion recognition for music therapy*, Journal on Multimodal User Interfaces, 5 (2012), pp. 27–35.

[14] S. Tripathi, A. Kumar, A. Ramesh, C. Singh, and P. Yenigalla, *Deep learning based emotion recognition system using speech features and transcriptions*, arXiv preprint arXiv:1906.05681, (2019).

[15] S. VE and Y. Cho, *Mrmr-eho-based feature selection algorithm for regression modelling*, Tehnički vjesnik, 30 (2023), pp. 574–583.

[16] X. Yang, Y. Dong, and J. Li, *Review of data features-based music emotion recognition methods*, Multimedia systems, 24 (2018), pp. 365–389.

[17] Q. Zhang, X. Chen, Q. Zhan, T. Yang, and S. Xia, *Respiration-based emotion recognition with deep learning*, Computers in Industry, 92 (2017), pp. 84–90.

[18] W. Zhou, J. Cheng, X. Lei, B. Benes, and N. Adamo, *Deep learning-based emotion recognition from real-time videos*, in Human-Computer Interaction. Multimodal and Natural Interaction: Thematic Area, HCI 2020, Held as Part of the 22nd International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part II 22, Springer, 2020, pp. 321–332.

Edited by: Rajanikanth Aluvalu
Special issue on: Evolutionary Computing for AI-Driven Security and Privacy: Advancing the state-of-the-art applications
Received: Jan 31, 2024
Accepted: Mar 12, 2024