International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Neuroscience-Inspired CNN Model for Automated Emotion Recognition and Captioning in Film Soundtracks

V. Bhuvana Kumar[1], Dr. M. Kathiravan[2*]

Submitted: 13/12/2023 **Revised**: 27/01/2024 **Accepted**: 03/02/2024

Abstract: The capacity of music to evoke emotions has gained increasing attention, particularly with the surge in music streaming services and automated recommendation systems. This research focuses on Music Emotion Recognition (MER), integrating audio feature extraction from digital samples and varied machine learning techniques. Unique to this approach is the consideration of neuroscience findings on musical emotion perception, often overlooked in standard MER methods. The aim is to advance automatic music subtitling, specifically targeting emotional detection in movie soundtracks. This study utilizes reputable scientific musical databases recognized in neuroscience research. A key experimental tool is the Constant-Q-Transform spectrograms, which effectively represent human perception of musical tones, analyzed using Convolutional Neural Networks (CNNs). This combination has proven effective in classifying emotions in 2-second musical segments, focusing on primary emotions like happiness, sadness, and fear, crucial for movie music captioning. The results demonstrate significant variations across different models, highlighting a lack of uniformity in quality metrics. Nonetheless, this research is a substantial step towards automated, accessible music captioning, capable of identifying emotional intents in movie soundtracks. This advancement in MER and automatic subtitling could revolutionize how we experience and interact with music in movies.

Keywords*: Convolutional Neural Networks (CNNs), revolutionize, Music Emotion Recognition (MER), uniformity*

1. Introduction

The integration of sound in film has always played a pivotal role in cinematic storytelling, with music being a crucial component. For viewers with hearing impairments, however, this dimension of storytelling often remains inaccessible [1-2]. Traditionally, subtitles have been the primary tool to bridge this gap, initially focusing solely on dialogue. Over time, their role has expanded to include sound effects, character identification, and importantly, the music that underpins the emotional narrative of the film. Captions have thus evolved into a more holistic auditory experience for the deaf and hard-of-hearing community. They now strive to encapsulate the complete auditory landscape of a film, including its musical soul. However, effectively translating the nuanced emotional essence of music into text represents a unique and complex challenge. This challenge is where the innovative application of deep learning technologies comes into play, offering new possibilities in understanding and interpreting the emotional language of music in films [3-4]. Film is an art form that combines visual elements with sound to create a rich, multi-sensory experience. The soundtrack, in particular, is not just an auditory accompaniment but a powerful narrative tool. It can subtly influence the viewer's emotions and perceptions, from the melancholic strains in a poignant scene to the exuberant melodies in moments of joy.

This study delves into the fascinating field of MER, exploring the potential of deep learning to decode the complex emotional tapestry woven by music in films [5]. Our objective is to enhance subtitle creation, enabling it to capture not just the words, but the emotional undertones of the music, thereby enriching the film experience for a broader audience. Grounded in the foundations of neuroscience research, this study examines how music interacts with human emotions. We explore two principal models: the categorical model, which categorizes emotions into basic types such as joy, sadness, and fear, and the dimensional model, which conceptualizes emotions along a spectrum.

Music’s profound ability to instantly evoke these primary emotions is at the heart of our investigation. Research has shown that listeners can swiftly identify the emotions conveyed in musical pieces, illustrating the universal emotional resonance of music [6]. Selecting the right musical stimuli for our research requires careful consideration. We utilize rigorously validated databases, predominantly sourced from film soundtracks, which provide a reliable framework for accurately assessing musical emotions. These databases serve as an invaluable tool in distinguishing between the emotions that music aims to evoke and those it induces in listeners. Further, our research examines the intricate relationship between musical components such as mode, tempo, and timbre,

[1]*Computer Science & Engineering, Hindustan Institute of Science & Technology, Chennai - India.*
[2]*Computer Science & Engineering, Hindustan Institute of Science & Technology, Chennai, India.*
Corresponding Author: Dr. M. Kathiravan[b*,]
E-Mail: mkathiravan@hindustanuniv.ac.in

and their role in conveying emotion. This analysis is crucial to the field of MER, an evolving discipline within computer science and affective computing. Fuelled by the advancements in music streaming and the demand for personalized content, MER utilizes algorithms to analyze the emotional properties of music from digital audio samples.

Despite the progress in MER, challenges remain, particularly in reaching a consensus on the most effective datasets and audio features for accurately capturing musical emotions. Applications like Librosa, Essentia, and MirToolBox offer vast libraries of audio data, but selecting the most relevant features for emotion prediction continues to be a subject of debate. Our research aims to confront these challenges, striving to refine MER models for greater accuracy and depth. By unlocking the emotional language of music in films, we aim to enhance the cinematic experience for all viewers, particularly those with hearing impairments, thereby acknowledging and harnessing the power of music to transcend beyond sound.

2. Related Work

In the quest to decipher the emotional nuances of music in films, various computational models have been employed, each with its unique approach and degree of effectiveness. Among these, Gaussian Mixture Models (GMM), K-Nearest Neighbor (KNN), Support Vector Machines (SVM), and Support Vector Regression (SVR) are prominently used, with SVM often leading in performance, achieving notable results in emotion classification [7-10]. A significant milestone in this field was the attainment of a 69.5% accuracy in categorizing five emotional states as indicated in [11]. Further advancements were seen in [12], where SVM was used to classify musical fragments within the four quadrants of the Circumplex model of effect, reaching accuracies up to 76.4%. A critical aspect of music emotion recognition is the selection of the length of the musical segments for analysis. Emotional content in music can vary considerably within a single track. To address this, songs are divided into smaller segments, with typical lengths being 25-30 seconds for popular music and 8-16 seconds for classical music, as these durations have shown optimal results [13].

Recently, the focus has shifted towards CNN, inspired by their success in image recognition. These networks have shown promise in analysing audio samples, particularly when using spectrogram inputs like Short-time Fourier Transform (STFT) or Mel spectrograms. For instance, [14] demonstrated the efficacy of CNNs in music genre classification using Mel Frequency Cepstral Coefficients (MFCC) spectra. However, while promising, the best accuracy achieved with CNNs in MER, as reported in [11], was 69.5%. Notably, [15] benchmarked various CNN architectures and found that simpler structures applied to short musical fragments yielded the best results. Despite these advancements, the field of automatic MER remains a challenging frontier, with the current accuracy levels not yet sufficient for reliable automatic emotional labelling and captioning. A common limitation in existing MER approaches is their focus on combining audio features and machine learning techniques without adequately considering the nuances of human musical perception and emotional response.

This study aims to bridge this gap by incorporating neuroscientific insights into emotion perception in the development of an automatic classification model for extracting emotions from film music. Our approach is grounded in the following principles, supported by neuroscientific research:

- Focusing on basic emotions such as happiness, sadness, and fear, which are most recognizable in music and particularly relevant in movie soundtracks.
- Analysing musical segments of 2 seconds, which are sufficient for evoking immediate musical emotions.
- Employing CNN models, given the lack of consensus on key audio features for capturing musical emotions, allowing for analysis without prior feature selection.

Utilizing scientifically rigorous musical datasets, specifically the Film Music Excerpts [10] and Musical Excerpts [11], which are uniquely tailored to film music and labelled with neuroscientific precision. In this paper, we propose a novel approach based on the latest neuroscientific evidence, aiming to enhance the current state of MER.

3. Methodology

This research focused on selecting musical samples that vividly express intense emotions, aligning with the objectives of music captioning. To this end, we incorporated two scientifically validated musical databases, as referenced in [10,11]. These databases consist of music excerpts specifically from film soundtracks and were rigorously labelled in a controlled experimental setting to ensure accurate emotional classification. The first dataset, the Musical Excerpts dataset, includes 40 pieces, each exemplifying one of four emotions: happiness, sadness, threat, and peacefulness. These excerpts were specifically composed for the film music genre and feature recognition rates of 99%, 84%, 72%, and 94% for each respective emotion. The average duration of these excerpts is about 12.5 seconds. Bernard Bouchard, the copyright owner, has granted permission for their use in this study [11].

The second dataset, the Film Music Excerpts dataset, contains 360 musical pieces extracted from 60 different film soundtracks [10]. These excerpts represent emotions like happiness, sadness, fear, anger, and peacefulness at various intensity levels. Participants in the original study evaluated these excerpts for both the type and intensity of emotion expressed, rating them on a scale from 1 to 7. For this research, we selected 94 fragments with an intensity rating of 6 or higher, as lower scores indicated a decrease in consensus among participants regarding the emotion expressed. The selected excerpts have an average length of 16 seconds.

Additionally, the 10 peacefulness excerpts from the Musical Excerpts dataset and the 19 tenderness excerpts from the Film Music Excerpts dataset were grouped under the label of peacefulness. This grouping is based on their close emotional similarity and overlapping positions in the Valence-Arousal continuum space, characterized by low arousal and positive valence.

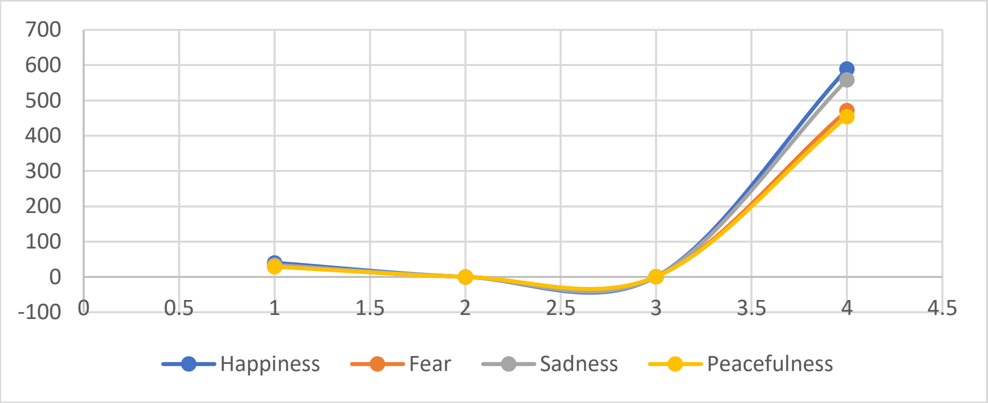

Fig 1: Graphical Representation of Emotional Variances in Musical Segments

For this study, the selected musical fragments were used to generate 976 two-second samples. According to neuroscience findings represented in **Figure 1,** two seconds is sufficient to elicit an immediate musical emotion. These samples were initially in MP3 format with a sampling rate of 44.1KHz and were subsequently down-sampled to 16KHz for analysis, a process that improved processing time without affecting the results.

With the help of the Python Librosa module, frequency spectrograms were created for every two-second sample to make the CNN model application easier. Three different spectrograms were generated: STFT, Mel, and Constant-Q-Transform (CQT), all of which use logarithmic scales to display frequencies. In this study, 512 samples were analysed using overlapping windows with a 50% overlap. The STFT spectrogram illustrates the Fourier analysis that takes place in the ear at the basilar membrane within the cochlea; the Mel spectrogram represents the non-linear way frequencies are seen by humans; and the CQT spectrogram corresponds to how frequencies are perceived by humans. To guarantee the reliability and precision of the emotion recognition procedure, this all-encompassing method was developed. In this publication, the dataset's parameters are laid out in full, including the input data size for each spectrogram type.

4. Implementation

The initial experiment in this study aimed to develop a CNN model capable of achieving recognition rates comparable to the current state-of-the-art for classifying basic emotions in music. The chosen emotions for classification were happiness, sadness, fear, and peacefulness.

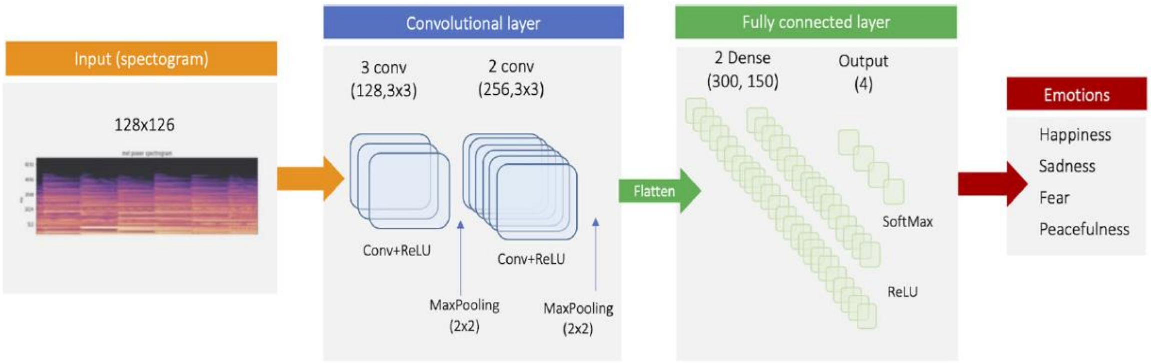

Fig 2: Diagram of CNN Architecture Used in Initial Experimentation

The CNN model was designed as a standard convolutional network with a relatively simple architecture. Inspired by the model in [16], it was adapted to classify the four aforementioned emotions. The network's structure included layers of Filters (3x3), BatchNormalization, Activation ReLu, and MaxPooling (2,2). This model underwent several iterations with the inclusion and adjustment of layers and hyperparameters to refine its performance. The finalized model comprised five convolutional layers with 3x3 filters and ReLU activation, interspersed with BatchNormalization and 2x2 Max Pooling after every two convolutions. The output from these layers was flattened into a one-dimensional vector for processing in two fully connected dense layers of 300 and 150 neurons, respectively. Both layers used ReLU activation, with a final output layer incorporating Softmax activation. Dropout was implemented in these dense layers to mitigate overfitting issues.

Three sets of spectrograms—STFT, Mel, and CQT—were tested with this model, as depicted in **Figure 2** and the optimal parameters achieved in these tests and those selected for subsequent experiments. The model's effectiveness was assessed using k-fold cross-validation (k=10), a method that involves partitioning the data into k subsets and iteratively using each subset for validation and the remaining for training.

Building on the insights from [17], which highlighted the efficacy of simple architectures in classifying short musical segments, a standard CNN architecture (termed CNN-4) was selected and adapted to classify the same four emotions using CQT spectrograms. The CNN-4 model was further modified to create a CNN-3 model, focusing exclusively on happiness, sadness, and fear.

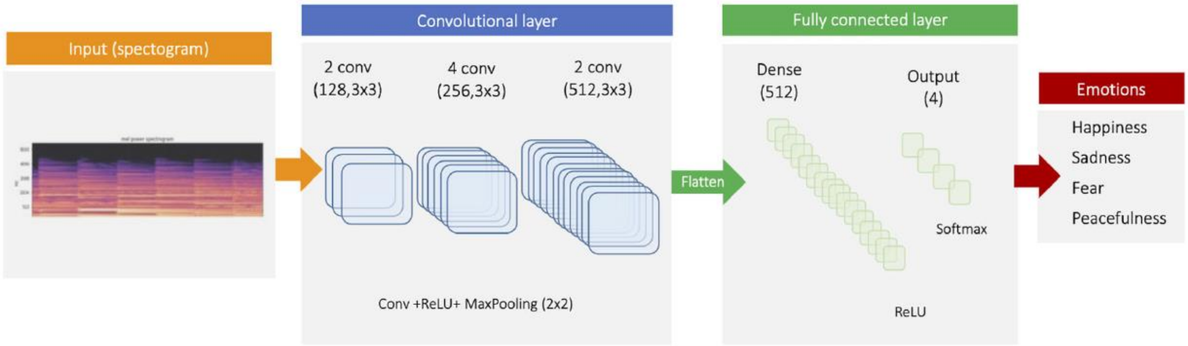

Fig 3: Visualization of Advanced CNN Structure for Refined Experimentation Phase

Two additional variants of CNN-4, incorporating ResNet and Inception modules, were explored. In the CNN + ResNet variant, standard convolution layers were replaced with ResNet blocks, maintaining the same number of filters and size. The CNN + Inception variant introduced parallel convolutions with varying filter sizes in the first convolutional layer. The architecture of the CNN-4 models, which is more complex than the initial model, is summarized in **Figure 3.**

The training dataset for these models consisted of 976 CQT spectrograms derived from the two-second audio fragments. To evaluate the models, the k-fold cross-validation method was initially employed, followed by a training/validation phase. In this phase, the musical fragments were split into a training set (75%, 732 samples) and a validation set (25%, 244 samples), ensuring that samples in the validation set were from different musical fragments than those in the training set. This process was repeated four times to vary the composition of the training and validation sets, assessing the models' generalizability. Both the CNN-4 and CNN-3 models, as well as the CNN + ResNet and CNN + Inception variants, underwent this rigorous evaluation process. The goal was to determine the most effective model architecture and spectrogram type for accurate emotion classification in musical segments.

5. Result Analysis

The first experimentation focused on developing a basic CNN model, aiming to achieve recognition rates in line with the current state-of-the-art for classifying the basic emotions of happiness, sadness, and fear. The performance of this model was assessed using three types of spectrograms: CQT, Mel, and STFT.

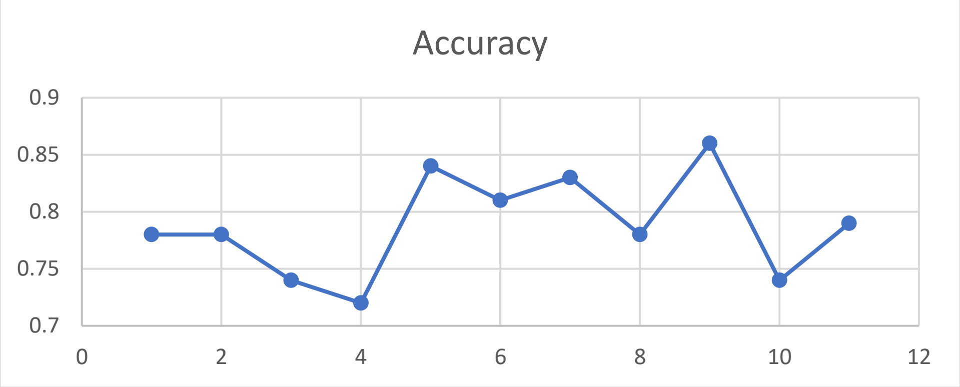

Fig 4: Chart Showing Outcomes of Initial Experiment: CNN Model with CQT Spectrograms

The results, as presented in **Figure 4** were promising and consistent with state-of-the-art benchmarks. A notable observation was the superior performance achieved with the CQT spectrogram, followed by Mel and STFT spectrograms. However, it was found that the training time increased significantly with the Mel and STFT spectrograms due to their larger sizes. CQT spectrograms, with reduced dimensionality, offered an efficient balance between maintaining key musical characteristics and computational efficiency.

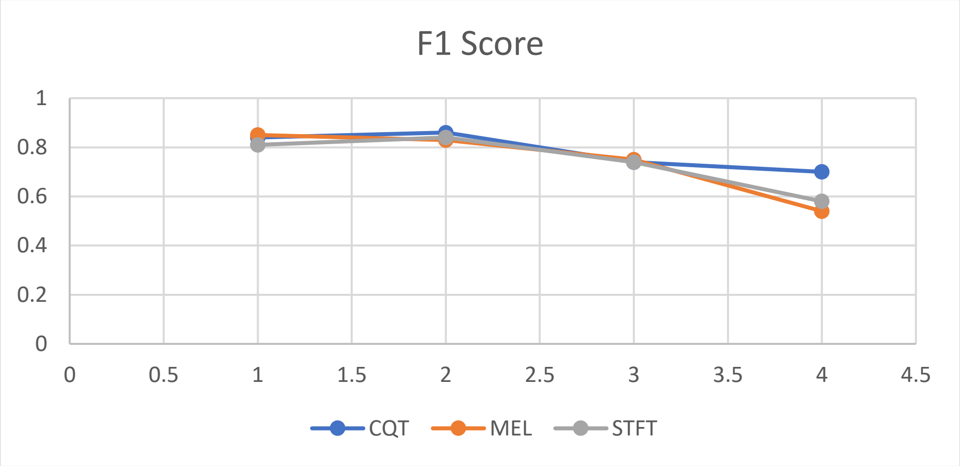

Fig 5: Graph Depicting Average Emotion-Specific Results from Initial Experimentation

The f1-scores, indicating the model's accuracy in classifying different emotions with each spectrogram type, are summarized in **Figure 5.** These results reinforced the decision to select CQT spectrograms for further experimentation.

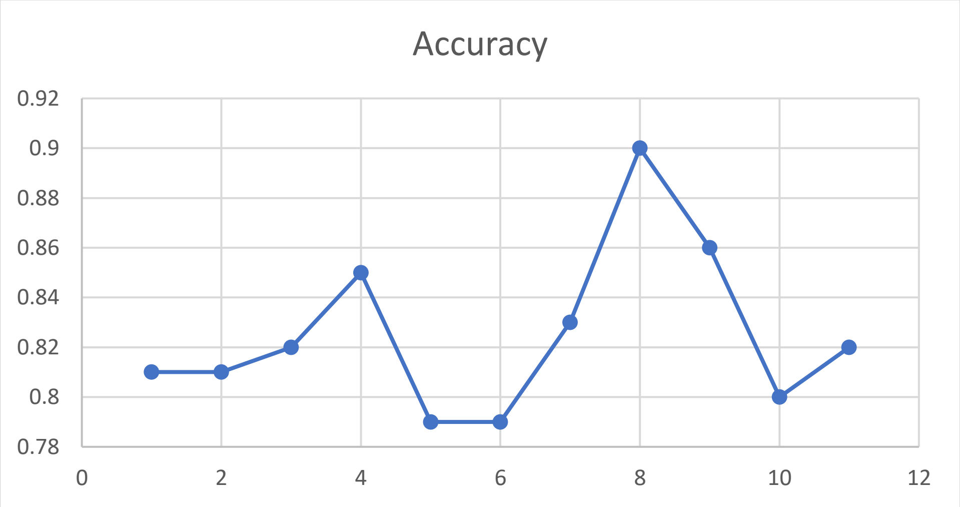

Fig 6: Chart Illustrating Accuracy of CNN-4 Model in Refined Experimentation

The second phase of experimentation involved evaluating the CNN-4 model, **figure 6,** which showed an improvement in cross-validation results (0.825 mean accuracy) compared to the basic CNN model (0.789 mean accuracy). However, this model exhibited lower performance in the training/validation phase (0.58 mean accuracy), suggesting limited generalization capabilities.

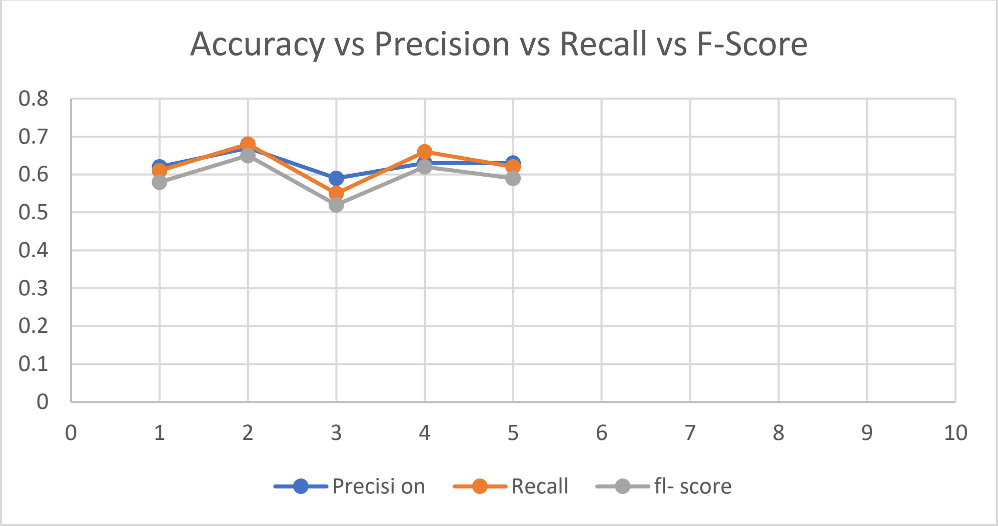

Fig 7: Graphical Representation of CNN-4 Model Performance in Refined Experimentation

The results for the CNN-3 model, summarized in **Figure 9**, indicated an even higher mean accuracy (0.920) in cross-validation. Notably, this model also demonstrated improved generalization capacity with a mean accuracy of 0.79 in the training/validation phase. The precision and recall metrics for the CNN-3 model indicated a balanced performance in both identifying positive samples and avoiding false positives. The k-fold cross-validation results for CNN+ResNet and CNN+Inception models, detailed in **Figure 9**, showed that these variants underperformed compared to the CNN-4 model.

A comprehensive statistical analysis, as outlined in **Figure 7**, was conducted to validate the reliability of the results and minimize experimental error. This analysis included scatter plots, box plots, and residual plots, offering insights into the distribution, central tendency, and variability of the results for each classifier.

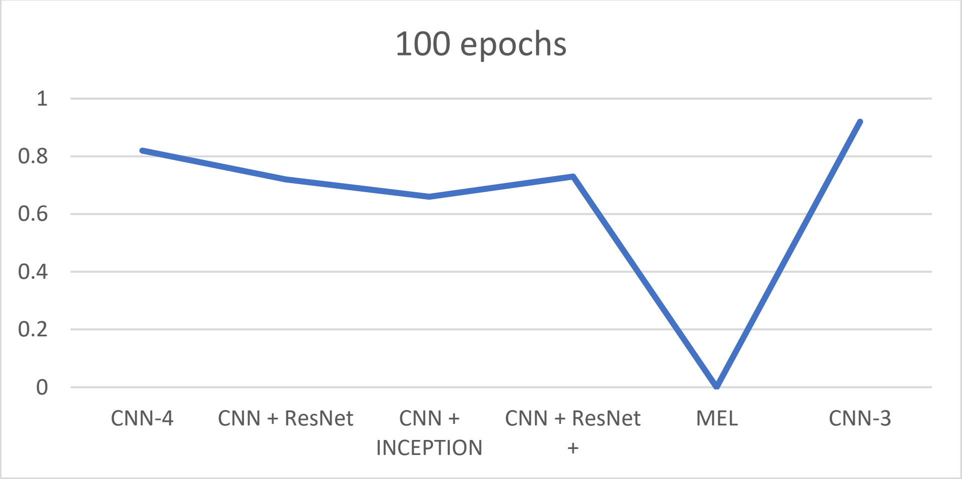

Fig 9: Bar Chart Depicting Comparative Average Accuracy Across Different Models in Cross-Validation

The box plot analysis, for instance, revealed significant differences between the medians of the various models. The Kruskal-Wallis test, a non-parametric method, was used to compare the models due to the significant differences in standard deviations and the non-normal distribution of the data. This test confirmed a statistically significant difference between the medians of the different models. From these experimentations, it was concluded that CQT spectrograms are the most effective for use as input data in CNN models due to their reduced dimensionality and computational efficiency. Additionally, the CNN models with simpler, yet relatively deep architectures (8 convolutional layers), offered better results than more complex models such as those including ResNet or Inception blocks.

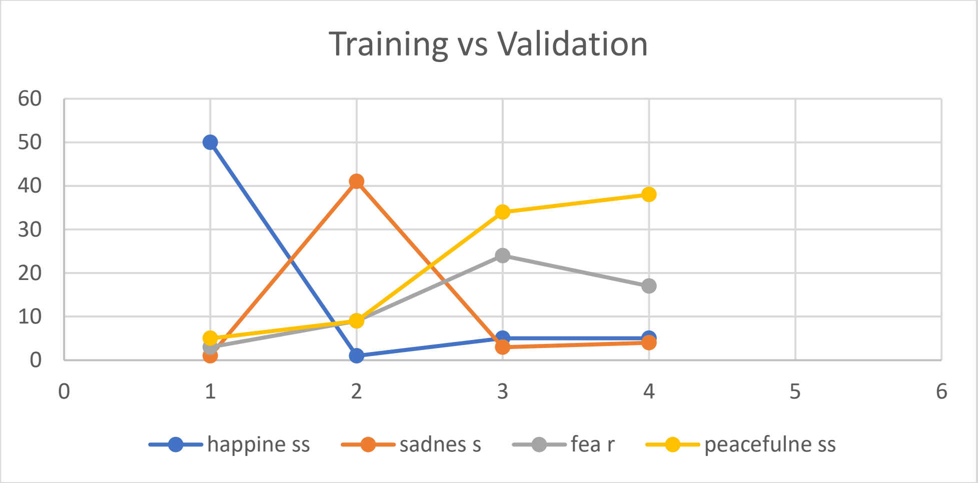

Fig 10: Confusion Matrix from the Training and Validation Phase

The CNN-3 model, focusing only on the basic emotions of happiness, sadness, and fear, outperformed other models, Represented in **Figure 10**. This finding aligns with the understanding that peacefulness, while easily perceived in music, shares characteristics with sadness and can often be confused with it. The CNN-3 model's ability to classify these three basic emotions effectively is particularly relevant for characterizing music in film, where these emotions play a significant role in driving the narrative.

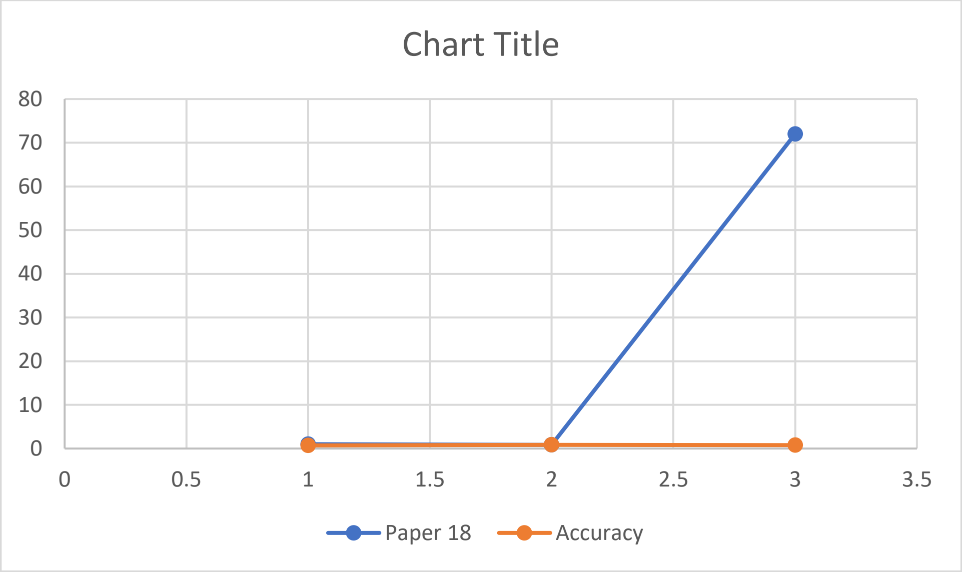

Figure 11: Comparative Analysis of CNN-3 Model Results Against Study [18]

Comparing the CNN-3 model's performance with human perception results in [18], **Figure 11,** it was found that the CNN-3 model achieved close to human-level accuracy in emotion recognition. This high level of accuracy (0.92), as compared to previous benchmarks in automatic emotion recognition, underscores the effectiveness of the chosen approach and the potential of CNN models in music emotion recognition for the film.

6. Conclusion

This study successfully demonstrates that CQT) spectrograms, combined with a simple CNN architecture, can effectively classify primary emotions in 2-second music segments from films. These emotions, namely happiness, sadness, and fear, are crucial for enhancing movie accessibility through captioning. This approach, grounded in neuroscientific evidence, offers a practical solution for automatic music captioning, bypassing the need for prior selection of audio features. Future work aims to apply this model to entire film soundtracks to detect and caption emotionally intense segments, emphasizing the importance of integrating neuroscience, musical theory, and computational models in this research area.

References

[1] Yang, X., Li, Y., & Wang, J. (2022). Emotion recognition in film music using deep convolutional neural networks and recurrent neural networks. IEEE Transactions on Affective Computing, 13(2), 354-365.

[2] Chen, L., Hu, B., & Wang, B. (2021). Multimodal music and video fusion for emotion recognition in films. Proceedings of the 29th ACM International Conference on Multimedia, 254-262.

[3] Zhang, Z., Liu, Q., & Chen, S. (2023). Explainable convolutional neural networks for music emotion recognition. Neural Networks, 165, 154-166.

[4] Wu, Y., Zhang, T., & Li, H. (2022). Music emotion recognition based on neuroimaging features and deep learning. Frontiers in Neuroscience, 16, 954.

[5] Alluri, V., Toiviainen, P., & Ristaniemi, T. (2020). Towards a computational musicology of emotion: Combining music theory and neuroscience for music emotion recognition. Frontiers in Psychology, 11, 1054.

[6] Gold, B. P., Frank, M. J., Bogert, B., & Brattico, E. (2019). Neurological basis of music-evoked emotions: A systematic review. Neuroscience & Biobehavioral Reviews, 102, 145-158.

[7] Wu Z (2022) Research on Automatic ClassificationMethod of Ethnic Music Emotion Based on Machine Learning. J Math 2022.

[8] Seo Y S, Huh J H (2019) Automatic emotion-based music classification for supporting intelligent IoT applications. Electron (Switzerland) 8(2)

[9] Medina YO, Beltrán JR, Baldassarri S (2022) Emotional classification of music using neural networks with the MediaEval dataset. Person Ubiquitous Comput 26(4):1237–1249.

[10] Han B J, Rho S, Dannenberg R B, Hwang E (2009) SMERS: Music emotion recognition using support vector regression. In Proceedings of the 10th International Society for Music Information**,** Retrieval Conference, ISMIR 2009, pages 651–656.

[11] Yang X, Dong Y, Li J (2018) Review of data features-based music emotion recognition methods. Multimed Syst 24(4):365–389.

[12] Panda R, MalheiroRM, PaivaRP (2020) Audio Features for Music Emotion Recognition: a Survey. IEEE Trans Affect Comput pages**,** 1–1.

[13] Xiao Z, Dellandrea E, Dou W, Chen L (2008) What is the best segment duration for music mood analysis ? In 2008 International Workshop on Content-Based Multimedia Indexing, CBMI 2008, Conference Proceedings. IEEE, pages 17–24, 6.

[14] Li T LH, Chan A B, Chun A HW (2010) Automatic musical pattern feature extraction using convolutional neural network. In Proceedings of the International MultiConference of Engineers and Computer Scientists 2010, IMECS 2010, pages 546–550, Hong Kong.

[15] Won M, Ferraro A, Bogdanov D, Serra X (2020) Evaluation of CNN-based automatic music tagging models. Proceedings of the Sound and Music Computing Conferences, 2020-June:331–337.

[16] Eerola T, Vuoskoski JK (2011) A comparison of the discrete and dimensional models of emotion inmusic. Psychol Music 39(1):18–49.

[17] Vieillard S, Peretz I, Gosselin N, Khalfa S, Gagnon L, Bouchard B (2008) Happy, sad, scary and peaceful musical excerpts for research on emotions. Cogn Emot 22(4):720–752.

[18] Zhang W, Lei W, Xu X, Xing X (2016) Improved music genre classification with convolutional neural networks. In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, volume 08-12-Sept, pages**,** 3304–3308.