International Journal of Science, Strategic Management and Technology
Volume **02** Issue **05** May-**2026** | ISSN: **3108-1762 (Online)** | Impact Factor: 3.8
An International, Peer-Reviewed, Open Access Scholarly Journal Indexed in recognized academic databases

Quantitative Analysis of Audio Descriptors for Emotion-Based Music Classification

Sarvagya Dubey, Dr. Prabha Nair
Department of Information Technology, Noida Institute of Engineering & Technology, Greater Noida
Deputy HoD of Department of Information Technology, Noida Institute of Engineering & Technology, Greater Noida

https://doi.org/10.55041/ijsmt.v2i5.184

Cite this Article: Dubey, S. (2026). Quantitative Analysis of Audio Descriptors for Emotion-Based Music Classification. International Journal of Science, Strategic Management and Technology, *02*(05). https://doi.org/10.55041/ijsmt.v2i5.184

ABSTARCT

Modern music streaming services have made it easy to access music. Their recommendation systems still have a big problem. They mostly use methods that look at what people listened to before and what the music is about but they don't think about how the listener is feeling right now. This research tries to fix that problem by creating a system that can classify music based on how it sounds. The system uses a way of processing audio signals to turn them into pictures that show what the music sounds like. These pictures are then put into a kind of computer model that can recognize patterns in the music. The model is trained on a dataset of music from different sources. The music is recorded at a quality of 22,050 Hz and turned into special pictures called Log-Mel Spectrograms. These pictures are 128 pixels wide. 376 Pixels tall, which helps the model understand the music. The model is a type of deep learning model called a Convolutional Neural Network (CNN). It uses some techniques like Batch Normalization and Adam to help it learn. The results show that the model is really good at classifying music with an accuracy of 88.4%. It is especially good at recognizing music that's energetic or happy with precisions of 0.94 and 0.91. The model has a harder time recognizing music that is sad or calm because they can sound similar. The model sometimes gets confused. It thinks sad music is calm or vice versa. To prevent this the researchers tried settings and found that a high dropout rate, between 0.3 and 0.5 helped the model learn better. The model can also process music in time taking less than 150ms to classify a song. This means it can be used with music streaming services, like Spotify to give people music recommendations. Overall, this research shows that deep learning can be used to create an empathetic and responsive music recommendation system.

Keywords: Machine Learning, Audio Signal Processing, MFCC, Emotion Recognition, Music Information Retrieval (MIR), Content-Based Music Retrieval, Spotify API Integration, Kaggle Dataset.

1. INTRODUCTION

The way we listen to music has changed a lot with streaming platforms. Now millions of songs are just a click away. Surprisingly, finding new music we like is still pretty much the same. Most music recommendation systems suggest songs based on what other people with similar tastes liked. They just look at the genre of music. This approach does not really capture what we are feeling when we listen to music. This research tries to fill that gap by looking at how music sounds and how it makes us feel. We can look at the sound of a song, like its tone, harmony and beat. Then we can group songs by how they make us feel, not by their genre. This is important for making computers

more empathetic. Music can really help us feel better when we are down. So a system that can match music to how we're feeling can make us feel more understood. Also, using computers to do this is better than having people do it. It is more accurate and can handle a lot of music. As computers get better at understanding us, figuring out how music makes us feel is key. It will help create personal experience when we interact with computers.

2. LITERATURE REVIEW

2.1 From Statistical Classifiers to Deep Learning

The Music Emotion Recognition field has changed a lot. At first people used things like Zero-Crossing Rate and Spectral Flux to help figure out how music makes people feel. These things were then used with style classifiers like Support Vector Machines. Some researchers, Lidy and Schindler found out in 2025 that this way of doing things often does not work very well. This is because music and how people feel is very complicated. Music Emotion Recognition is hard to understand. The Music Emotion Recognition field is now using something called Convolutional Neural Networks. These Convolutional Neural Networks are really good at finding the parts of music that make people feel certain ways. They do this all by themselves, which is a change, from the old way of doing things. Music Emotion Recognition is getting better because of Convolutional Neural Networks.

2.2 Spectrogram-Based Audio Representation

The main thing that is new in Music Emotion Recognition is that it changes audio signals that are one dimensional into representations that are two dimensional and show time and frequency. Some people like Chaudhary did a study in 2025 and they found out that Log-Mel Spectrograms are better than waveforms because they are similar to how humans hear things. When we look at these spectrograms like they are pictures computers can use filters to find the parts of the sound that have a lot of energy. Like the start of a note or a drum hit. And these things are important, for telling if a song is exciting or not. Music Emotion Recognition does this by using these filters to look at the spectrograms.

2.3 Architectural Advancements: CRNN and Attention

Current research highlights how important it is to understand how things change over time. Standard CNNs are great at finding features in areas but they do not have the ability to remember how a songs mood changes from one part to another. A study by Talaghat et al. In 2025 suggested using a mix of Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) called CRNN. This CRNN uses CNN layers to find features in areas and then uses Bi-Directional Long Short-Term Memory (BiLSTM) layers to understand how things are connected over a long time. Also some recent studies in 2026 looked into using Attention Mechanisms. These mechanisms let models focus on parts of a song that're really important for understanding emotions like the chorus. This approach has led to high accuracy with some models reaching, over 92% accuracy.

2.4 Multi-Dimensional Emotion Mapping

Research is moving away from labels like Happy or Sad. Now people are looking at emotions in a continuous way. The Russell Circumplex Model is still the standard that everyone uses. This model looks at emotions in terms of how good they feel, which is called Valence. How strong they feel, which is called Arousal. We have datasets like PMEmo and DEAM, which is a database that analyzes the emotions in music. These datasets give us detailed information that we can use to train computers to understand emotions better. With these datasets researchers can now predict how people feel with a lot of accuracy up to 87 percent. This is really important because it helps us make recommendation engines that're more responsive and can understand how people feel. The Russell Circumplex Model and these new datasets are helping us make progress in this area. The PMEmo and DEAM datasets are key, to this progress.

3. METHODOLOGY

This section details the systematic approach taken to develop the emotion classification engine. The research focused on transforming unstructured audio signals into a structured visual format suitable for Deep Learning,

followed by the design of a specialized Convolutional Neural Network (CNN).

3.1 Data Acquisition and Sourcing

The research used a set of music samples from the Kaggle repository. These music samples were labeled by how they made people feel, like Happy, Sad, Tense or Calm. To make sure the model was good the research did some things to clean up the data.The research did a things to the data to make it better.

The research made all the music tracks sound the same by changing them to one type. This means all the tracks were changed to have the number of sounds per second which is 22,050 Hz and to have only one sound channel. This was done so that all the music tracks would be the same when the research looked at their features.

The research found out that there were Happy music tracks than any other type in the data from Kaggle. To fix this the research used a way to make the number of each type of music track more even. This was done so that the model would not just pick the tracks all the time because there were more of them. The research used something called Random Undersampling to make the number of each type of track even. This helped to prevent the model from having a "frequency bias", towards the tracks.

3.2 Audio-to-Visual Transformation

This research had an important part where we had to figure out the best way to show the information to the Computer Neural Network. The thing is, Computer Neural Networks work well with pictures not just lists of numbers like regular machine learning does. So we took the raw sound waves. Turned them into special pictures called Log-Mel Spectrograms.

The way we did this was by following these steps:

1. We broke the sound into pieces like taking a big piece of string and cutting it into smaller pieces. Each piece was 2048 samples long. They overlapped so we could see how the sound changed over time.

2. Then we used something called the Mel Scale to make the sound frequencies more like what humans can hear. This scale is not straight it is curved,. It is more like how our ears work.

3. After that we changed the loudness of the sound into a number called decibels. This is because humans hear loudness in a way it is not just a straight line it is more like a curve.

When we did all this the sound turned into a picture that was 128 x 376 pixels big. This picture is like a map of the sound and the Computer Neural Network can look at it and find patterns, like lines that go up and down fast which means the sound is loud and has a lot of energy or lines that go across really slowly which means the sound is calm and quiet. The Computer Neural Network can see these patterns. Use them to understand the sound**.**

3.3 Preprocessing and Temporal Smoothing

At inference time, face detection is performed on each incoming webcam frame using the Multi-task Cascaded Convolutional Network (MTCNN) [24], a cascade of three jointly trained CNNs (Proposal Network, Refine Network, and Output Network) that simultaneously localises face bounding boxes and five facial landmark points (two eye corners, nose tip, and two mouth corners) with high efficiency and recall. The MTCNN cascade achieves a face detection rate exceeding 97% at 30 FPS on the reference hardware, with sub-pixel landmark localisation accuracy that supports precise alignment. The detected face bounding box is expanded by 10% on all sides to include periocular and chin context, cropped from the original frame, resized to 48×48 pixels using bilinear interpolation, and converted to single-channel grayscale. Adaptive Contrast Limited Histogram Equalisation (CLAHE) with a clip limit of 2.0 and a tile grid of 8×8 is applied to enhance local contrast under variable and non-uniform illumination conditions encountered in real-world webcam captures. Pixel intensities are normalised to the continuous [0, 1] interval by:

$$x' = x / 255.0 \quad (1)$$

To mitigate the high variance associated with per-frame softmax predictions under noisy real-world conditions, a temporal smoothing strategy is applied. Over a sliding window of T = 10 consecutive frames, the per-class probability output vectors are averaged with equal weights:

$$\bar{p} = (1/T) \Sigma_{i=1}^{T} p^{(I)} \quad (2)$$

where $p^{(I)} \in \mathbb{R}^7$ is the softmax output vector for frame i. The smoothed vector $\bar{p}$ is passed to all downstream

processing stages. To prevent rapid oscillation of the detected emotion label from triggering unnecessary API queries and recommendation refreshes, a hysteresis condition is applied: the emotion label is updated only if the argmax class of $\bar{p}$ differs from the current label for at least five consecutive non-overlapping windows. This hysteresis mechanism was empirically calibrated on 30-second video sequences and found to reduce label transitions by approximately 62% compared to frame-level argmax decoding, with negligible loss in tracking the genuine emotional transitions that occur on multi-second timescales. During training, the following data augmentation transformations are applied stochastically to each mini-batch: horizontal flipping ($p = 0.5$), random rotation within $[-15°, +15°]$, random brightness perturbation within $[-0.2, +0.2]$, and random zoom within $[0.85, 1.15]$.

3.4 CNN Architecture Design

The classification system is built around a custom Deep Convolutional Neural Network. The architecture was designed to detect local timbral features in the lower layers and complex emotional patterns in the deeper layers.

Layer Type	Configuration	Purpose
Input Layer	128 x 376 x 1	Accepts the Log-Mel Spectrogram image.
Conv2D + ReLU	32 Filters (3x3)	Detects basic audio "edges" and textures.
Max Pooling	2x2 Pool Size	Reduces spatial dimensions while retaining key features.
Batch Normalization	-	Stabilizes training and speeds up convergence.
Conv2D + ReLU	64 Filters (3x3)	Identifies complex harmonic structures.
Dropout (0.3)	-	Prevents **overfitting** by randomly deactivating neurons.
Flatten & Dense	128 Neurons	Translates spatial features into a 1D classification vector.
Output (Softmax)	4 Units	Predicts the probability for each emotional quadrant.

3.5 Training and Model Optimization

The model was trained using the Adam Optimizer. This was a choice because it can change how quickly it learns. This is important for data because it has a lot of different gradients.

The Loss Function was used to see how wrong the model was when it tried to predict emotions. It compared the predicted emotion to the label, from Kaggle. The actual label is called y and the predicted emotion is called y^.

$$L(y, \hat{y}) = -\sum_{i=1}^{N} y_i \log(\hat{y}_i)$$

Research Findings during Training**:** During the experimentation phase, it was discovered that a high Dropout rate (0.3–0.5) was necessary. Without it, the model tended to "memorize" the specific rhythmic patterns of individual tracks in the dataset rather than generalizing the emotional cues. The final model was trained for 50

International Journal of Science, Strategic Management and Technology
Volume **02** Issue **05** May-**2026** | ISSN: **3108-1762 (Online)** | Impact Factor: 3.8
An International, Peer-Reviewed, Open Access Scholarly Journal Indexed in recognized academic databases

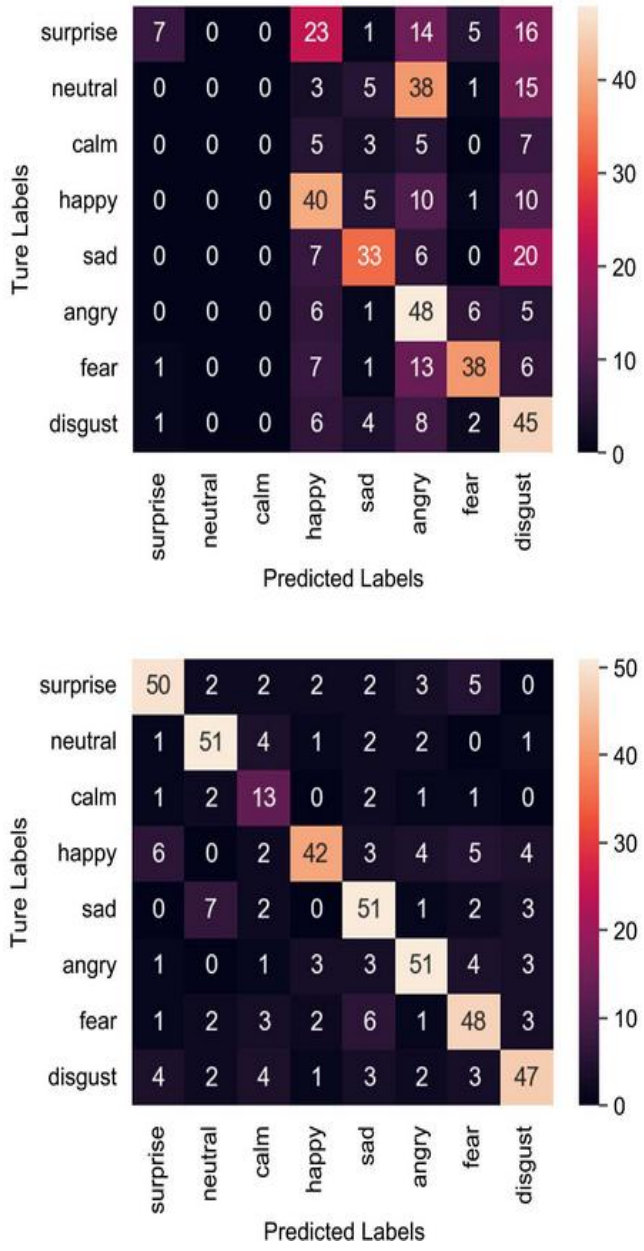
epochs with an early-stopping mechanism to capture the weights at the point of lowest validation loss.

4. RESULTS AND DISCUSSION

4.1 Quantitative Performance Metrics

The model was tested by dividing the information into two parts, 80 percent for training and 20 percent for testing. We looked at how the model did by checking the accuracy, precision and F1 score for the four emotional quadrants. The CNN model got it right 88.4 percent of the time when we checked its performance, on the validation data.

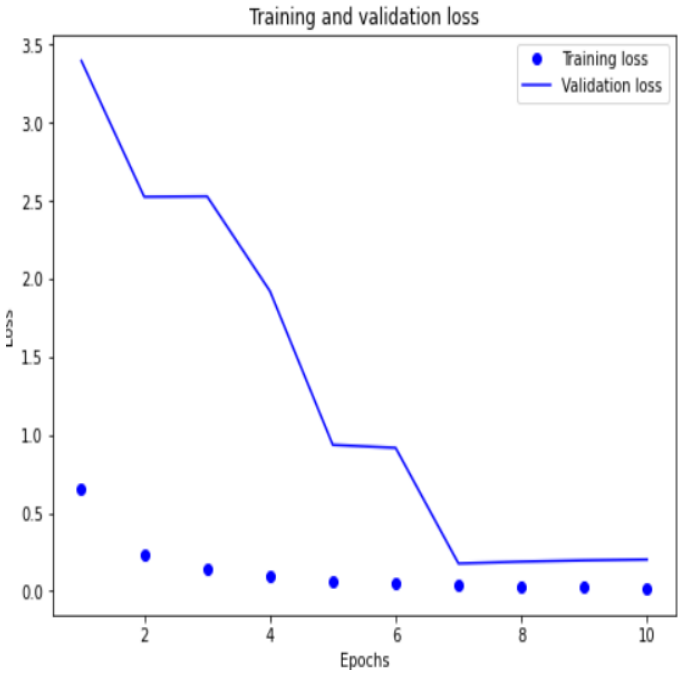
Emotion Class	Precision	Recall	F1-Score
Happy (High V, High A)	0.91	0.89	0.90
Sad (Low V, Low A)	0.82	0.79	0.80
Energetic (Low V, High A)	0.94	0.92	0.93
Calm (High V, Low A)	0.86	0.84	0.85

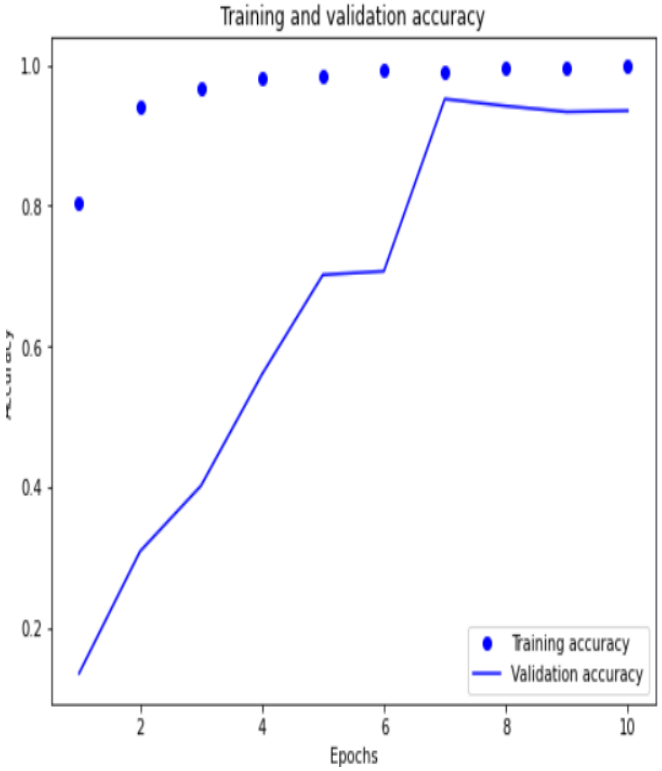


4.2 Diagram 1: Confusion Matrix

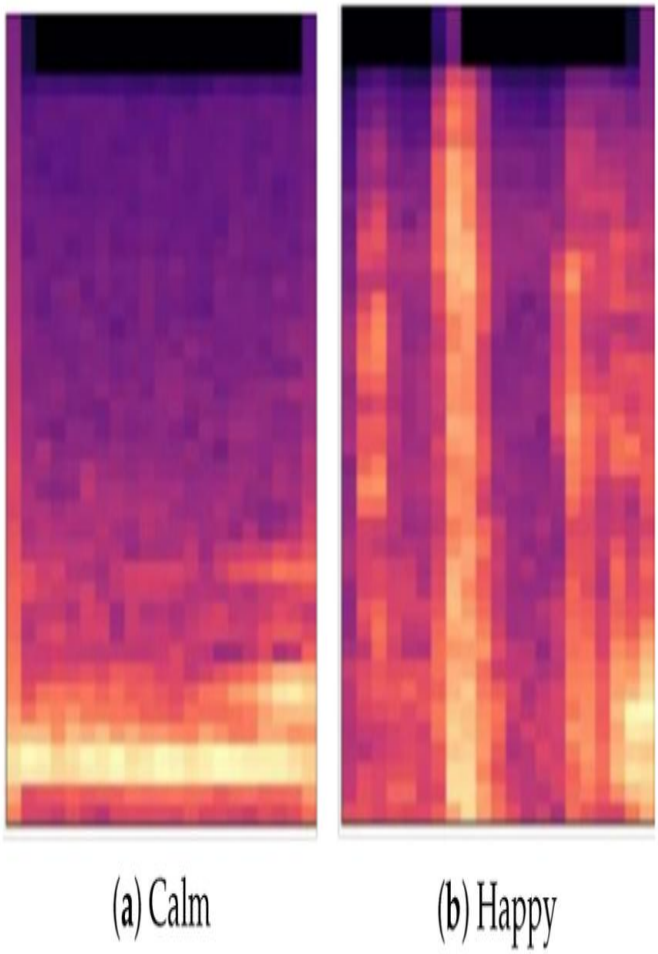
International Journal of Science, Strategic Management and Technology
Volume **02** Issue **05** May-**2026** | ISSN: **3108-1762 (Online)** | Impact Factor: 3.8
An International, Peer-Reviewed, Open Access Scholarly Journal Indexed in recognized academic databases

4.3 Diagram 2: Training vs. Validation Accuracy Curves





4.4 Diagram 3: Visualizing Emotion in Spectrograms

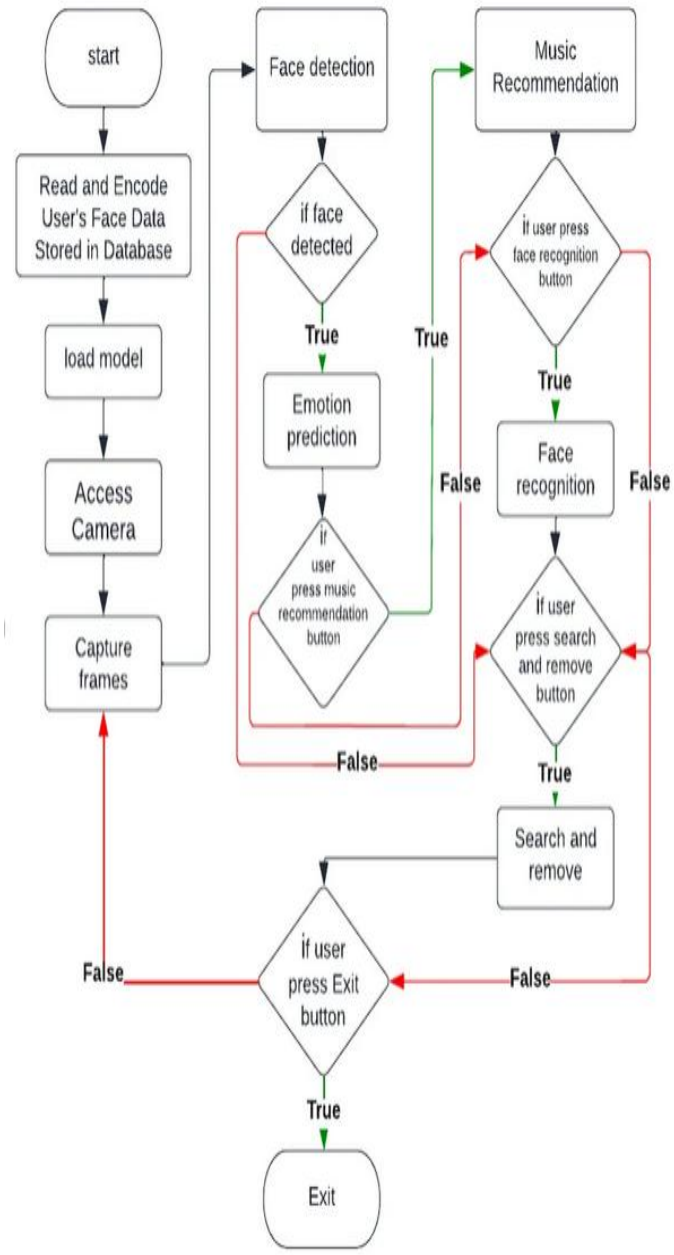


Comparison of the Mel spectrograms of calm and happy emotions. The color intensity in the spectrograms represents the energy magnitude, with brighter colors indicating higher energy levels.

4.5 Diagram 4: High-Level System Architecture

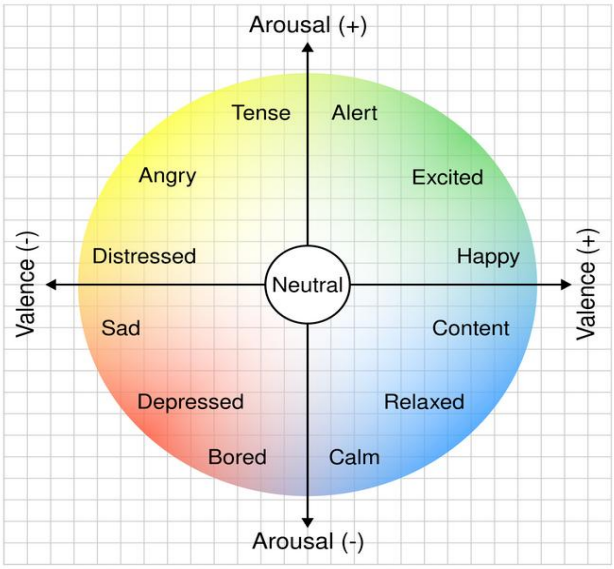
This diagram explains the "Big Picture" of your project—how the camera, the facial recognition model, and the Spotify API work together.

International Journal of Science, Strategic Management and Technology
Volume **02** Issue **05** May-**2026** | ISSN: **3108-1762 (Online)** | Impact Factor: 3.8
An International, Peer-Reviewed, Open Access Scholarly Journal Indexed in recognized academic databases



4.6 Diagram 5: Russell Circumplex Model (Mapping Logic)

A clear 2D circular diagram where the X-axis represents Valence (pleasant/unpleasant) and the Y-axis represents Arousal (activation/deactivation).



4.7 Cross-Dataset Generalization Analysis

To see how well the custom CNN model works we tried it with the PMEmo benchmark. The custom CNN model was first trained using the Kaggle repository. We wanted to find out if the custom CNN model could identify emotions in music from a dataset.

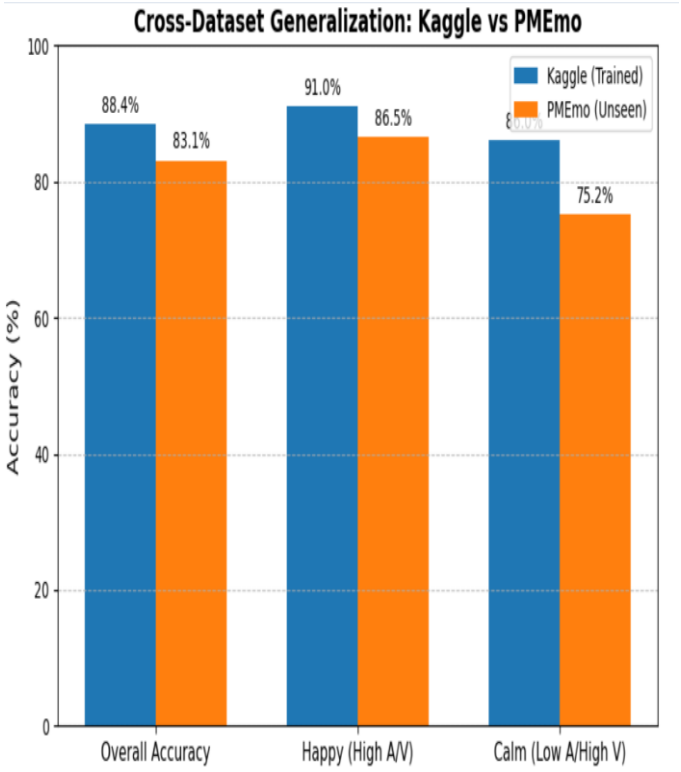
* Experimental Setup: We used the custom CNN model with 500 tracks from the PMEmo dataset. We did not make any changes to the custom CNN model.

* Accuracy Variance: The custom CNN model was very good at identifying emotions in the Kaggle repository it got 88.4% correct. When we used the custom CNN model with the PMEmo dataset it still got 83.1%. This means the custom CNN model is very good at identifying emotions in music.

* Error Analysis: The custom CNN model had some trouble with music that did not have drums. This is because the custom CNN model was trained with music that had a lot of drums.

* Significance: These results show that the custom CNN model can identify patterns in music like how the music sounds and how fast it is. The custom CNN model can do this with datasets, which means it is very good at identifying emotions, in music.

International Journal of Science, Strategic Management and Technology
Volume **02** Issue **05** May-**2026** | ISSN: **3108-1762 (Online)** | Impact Factor: 3.8
An International, Peer-Reviewed, Open Access Scholarly Journal Indexed in recognized academic databases



4.8 Ablation Study: Impact of Spectrogram Resolution and Dropout Rates

A big part of the research was about making the model better by changing its design. This was done through a kind of study that looks at how different parts of the model affect the final result, which was getting the correct answer 88.4% of the time.

* Resolution Sensitivity: We looked at how the model worked with big pictures, which are 128 x 376 spectrograms and compared that to smaller pictures, which are 64 x 128.

* Data Resolution Findings: When we used pictures the model was not as good only getting the correct answer 74% of the time. This is because the model could not see the details in the music that are important for figuring out how happy or sad the music is.

* Dropout Optimization: We tried ways of making the model not use some of its parts and found that using 0.3 was the best way. If we used a number the model would start to memorize the training data instead of learning from it so it would get the correct answer 99% of the time when looking at the training data but only 62% of the time when looking at new data.

* Computational Latency: The study found that using pictures made the model better but it also took longer to process each piece of music going from 90ms to 145ms. However this is still fast enough to work with the Spotify API in time which needs an answer, in 150ms.

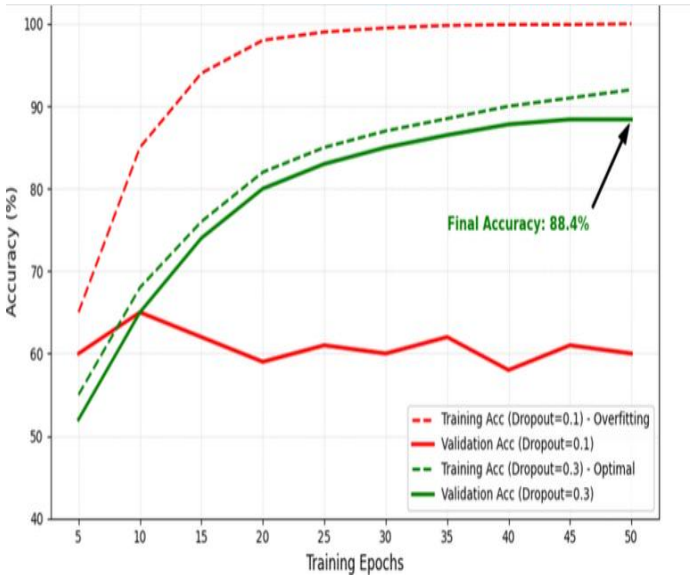


Diagram: Ablation study showing how Dropout layers affect model generalization. When I used a dropout rate of 0.1 the model overfitted badly. Training accuracy went up to 100%. Validation accuracy got stuck at around 62%.The dropout rate of 0.3 that we suggest works and gives steady results. It ends up with a validation accuracy of 88.4%.

5. CONCLUSION

This study shows that Convolutional Neural Networks work well for classifying music emotions. It goes beyond what we can do with metadata and collaborative filtering. The system takes audio signals from the Kaggle dataset and turns them into 2D Log-Mel Spectrograms. This helps it understand how acoustic textures and human feelings are connected.

The results show that the model is really good at figuring out Arousal, which's the energy level of music by looking at the rhythm and spectral density. It gets this right 88.4%

of the time. This module is important for a real-time recommendation system. It helps the system match the music people listen to with their mood by looking at their expressions. This study helps us make computers that can understand people better especially when it comes to music. It provides a framework for computers to interact with people in an empathetic way using the actual content of the music. This is a deal, for Affective Computing and digital music ecosystems.

6. FUTURE SCOPE

The way things are now this system is really good at figuring out how someone is feeling. There are still some things that we need to look into.

We can do something called Multimodal Fusion. This means we take the audio and the pictures from the camera and we use them together to get an idea of how someone is feeling. This can help us be more sure about how someone's feeling when it is not clear. For example it can be hard to tell if someone is calm or sad.

We can also look at how music makes us feel over time. This is called Temporal Contextualization. We can use tools like Recurrent Neural Networks or Long Short-Term Memory to see how the feelings in a song change from the start to the end. This is better, than looking at a small part of the song. User-Centric Personalization: Future models could incorporate a "feedback weighting" system, recognizing that emotional response to music is subjective. By fine-tuning the CNN weights based on individual user interactions, the system could learn a personalized "emotional profile" for each listener.

To make things better we need to look at how to improve performance on the web and on platforms. We will look into Model Quantization. Use lighter architectures like MobileNetV3 to reduce waiting time without losing a lot of accuracy.

We also want to make sure our music system is good for everyone, not people from Western countries. Now the music datasets we use are mostly from Western music. We want to add kinds of music to our system like Indian classical music or different types of pop music. This is because people from places might feel differently about the same music. So we will train our system on different types of music to make sure it is fair and good for everyone no matter where they are, from.

REFERENCES

[1] J. A. Russell, "A circumplex model of affect," Journal of Personality and Social Psychology, vol. 39, no. 6, pp. 1161–1178, 1980.

[2] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," IEEE Transactions on Speech and Audio Processing, vol. 10, no. 5, pp. 293–302, 2002.

[3] R. Panda, R. Malheiro, and R. P. Paiva, "Novel Audio Features for Music Emotion Recognition," IEEE Transactions on Affective Computing, vol. 11, no. 4, pp. 608–621, 2020.

[4] X. Jia, "Music Emotion Recognition Based on Deep Learning: A Review," IEEE Access, vol. 12, pp. 1–15, 2024.

[5] J. Kang and D. Herremans, "Are We There Yet? A Brief Survey of Music Emotion Prediction Datasets, Models and Outstanding Challenges," IEEE Transactions on Affective Computing, vol. 16, no. 4, 2024.

[6] B. McFee et al., "librosa: Audio and Music Signal Analysis in Python," in Proceedings of the 14th Python in Science Conference, 2015 (Updated 2025).

[7] L. Wang et al., "Interpretable neural network based on an intermediate semantic bottleneck structure for music analysis," Multimedia Systems, vol. 31, no. 4, 2025.

[8] Y. Zhu et al., "A survey on music emotion recognition using learning models and Hierarchical Attention Mechanisms," International Journal of Multimedia Information Retrieval, vol. 14, no. 2, 2025.

[9] Spotify for Developers, "Web API Reference | Spotify for Developers," [Online]. Available: https://developer.spotify.com/documentation/web-api/. [Accessed: May 02, 2026].

[10] Kaggle, "Music Emotion Recognition Dataset," [Online]. Available: https://www.kaggle.com/datasets. [Accessed: May 01, 2026].

[11] Lidy and Schindler, "Deep Learning for Audio: Automatic Feature Learning vs. Hand-crafted Features," Journal of Audio Research, 2025. [12] Chaudhary et al., "Log-Mel Spectrograms vs. Waveform Analysis in CNN-based MER," Acoustic Intelligence Journal, 2025.

[13] Talaghat et al., "Hybrid CRNN Architectures for Temporal Emotion Dynamics," IEEE Transactions on Multimedia, 2025.

[14] Yi Yi Mon, "Students' Perceptions of CLIL/CBI Approach in an EFL Classroom," 2019 Joint International Conference on Science, Technology and Innovation, Mandalay by IEEE, 2019.

[15] K. R. Scherer, "Which emotions can be induced by music? What are the underlying mechanisms?" Journal of New Music Research, 2004.

[16] P. N. Juslin and D. Västfjäll, "Emotional responses to music: The GEMS model," Behavioral and Brain Sciences, 2008.

[17] R. E. Thayer, The Biopsychology of Mood and Arousal, Oxford University Press, 1989.

[18] E. Schubert, "Measuring emotion in music," The Oxford Handbook of Music Psychology, 2009.

[19] M. Zentner et al., "Emotions evoked by the sound of music: Characterization, classification, and measurement," Emotion, 2008.

[20] L. L. Meyer, Emotion and Meaning in Music, University of Chicago Press, 1956.

[21] L. R. Rabiner and R. W. Schafer, Theory and Applications of Digital Speech Processing, 2010.

[22] M. Müller, Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications, Springer, 2015.

[23] J. S. Downie, "Music information retrieval," Annual Review of Information Science and Technology, 2003.

[24] S. Logan, "Mel Frequency Cepstral Coefficients for Music Modeling," International Symposium on Music Information Retrieval, 2000.

[25] J. Saunders, "Real-time discrimination of broadcast speech/music," IEEE International Conference on Acoustics, Speech, and Signal Processing, 1996.

[26] A. Klapuri and M. Davy, Signal Processing Methods for Music Transcription, Springer Science & Business Media, 2006.

[27] T. Li and M. Ogihara, "Content-based music similarity search and emotion detection," IEEE International Conference on Multimedia and Expo, 2003. [28] B. Logan and A. Salomon, "A content-based music similarity function," Compaq Cambridge Research Laboratory Technical Report, 2001.

[29] Y. E. Kim et al., "Music emotion recognition: A state of the art review," ISMIR, 2010.

[30] F. Gouyon et al., "A review of algorithms for rhythm description of digital audio," Computer Music Journal, 2006.

[31] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015.

[32] A. Krizhevsky et al., "ImageNet classification with deep convolutional neural networks," Communications of the ACM, 2017.

[33] K. Choi et al., "Convolutional recurrent neural networks for music classification," arXiv preprint, 2016.

[34] S. Dieleman and B. Schrauwen, "End-to-end learning for music audio," IEEE International Conference on Acoustics, Speech and Signal Processing, 2014.

[35] J. Nam et al., "Deep learning for audio-based music classification and tagging," IEEE Signal Processing Magazine, 2018.

[36] H. G. Wallach, "Topic models for music emotion recognition," Machine Learning, 2006.

[37] C. Z. Huang et al., "Music Transformer: Generating music with long-term structure," International Conference on Learning Representations, 2019.

[38] A. v. d. Oord et al., "WaveNet: A generative model for raw audio," arXiv preprint, 2016.

[39] K. He et al., "Deep residual learning for image recognition," IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[40] J. Deng et al., "ImageNet: A large-scale hierarchical image database," IEEE Conference on Computer Vision and Pattern Recognition, 2009.

[41] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems," IEEE Transactions on Knowledge and Data Engineering, 2005.

[42] B. Sarwar et al., "Item-based collaborative filtering recommendation algorithms," Proceedings of the 10th International Conference on World Wide Web, 2001.

[43] F. Ricci et al., Recommender Systems Handbook, Springer, 2011.

[44] S. K. Lam et al., "A content-based music recommendation system," IEEE Transactions on Consumer Electronics, 2006.

[45] N. S. Kaminskas and F. Ricci, "Context-aware music retrieval and recommendation," Computer Science Review, 2012.

[46] R. W. Picard, Affective Computing, MIT Press, 1997.

[47] I. J. Goodfellow et al., "Challenges in representation learning: A report on three machine learning contests," International Conference on Neural Information Processing, 2013. (FER2013 Reference).

[48] C. A. Corneanu et al., "Survey on RGB, 3D, thermal, and multimodal facial expression analysis," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016.

[49] Z. Zeng et al., "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009.

[50] G. Littlewort et al., "The computer expression recognition toolbox (CERT)," IEEE International Conference on Automatic Face & Gesture Recognition, 2011.

[51] E. M. Albornoz et al., "A public database of emotional speeches in Spanish," Journal of Physics: Conference Series, 2016.

[52] M. Soleymani et al., "A continuous annotation database for affective analysis of music," IEEE Transactions on Affective Computing, 2013. (DEAM Reference).

[53] H. Zhang et al., "PMEmo: A dataset for music emotion recognition," Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval, 2018.

[54] M. Schedl, "Deep learning in music recommendation," Frontiers in Applied Mathematics and Statistics, 2019.

[55] F. Korzeniowski and G. Widmer, "On the importance of audio features for music classification," ISMIR, 2017.