ISSN: 2230-9926

Available online at http://www.journalijdr.com

IJDR

International Journal of Development Research
Vol. 16, Issue, 05, pp. 70422-70426, May, 2026
https://doi.org/10.37118/ijdr.30821.05.2026

RESEARCH ARTICLE **OPEN ACCESS**

WISSH – WALLPAPER INNOVATOR BY SYNTHESIZING SONG HALLMARK

***Prof. Lalita Panika, Amit Kumar Jha, Raj Sahu, Aryan Prithwani and Siddhant Shukla**

Bhilai Institute of Technology, Department of Computer Science Engineering Kendri, Raipur, India

ARTICLE INFO

Article History:

Received 11th February, 2026
Received in revised form
26th March, 2026
Accepted 17th April, 2026
Published online 25th May, 2026

Key Words:

Social media, Student Learning, Learning Motivation, Education management, Human Resource management.

****Corresponding author: Prof. Lalita Panika***

ABSTRACT

This paper presents an intelligent system that transformsaudio music into visually expressive animated wallpapers.Leveraging deep learning-based mood recognition and advancedimage generation via Stable Diffusion, the project combinesacoustic feature extraction with a culturally aware mood classifierto interpret musical emotions. These emotions are converted intostylistic prompts that guide the creation of aesthetic, animatedwallpapers. The system integrates PyTorch for mood detection,HuggingFace Transformers for audio embeddings, and Diffusersfor visual generation. The platform offers both single and batchprocessing, enabling immersive visual experiences from soundinput. By bridging the gap between auditory emotion and visualstorytelling, this project opens a pathway for more immersivemultimedia experiences. The paper explores both the algorithmicdesign and user-centric features of the system, demonstratingpotential applications in art therapy,music streaming platforms, and personalized device aesthetics.

Citation: Prof. Lalita Panika, Amit Kumar Jha, Raj Sahu, Aryan Prithwani and Siddhant Shukla, 2026. *"*WISSH – Wallpaper Innovator by Synthesizing Song Hallmark". *International Journal of Development Research*, 16, (05), 70422-70426.

INTRODUCTION

The emotional power of music has long been recognizedacross cultures, serving as a universal language that conveys feelings, memories, and atmospheres. In recent years, researchers and developers have explored the intersection ofsound and visuals, aiming to create more immersive and emotionally engaging multimedia experiences. Translating audio based moods into visual art not only enriches aesthetic expression but also opens novel opportunities in human-computer interaction, media personalization, and therapeutic applications. This paper introduces a next-generation AI-powered applicationthat analyzes the emotional content of music and generatescorresponding animated wallpapers. The system bridgesthe gapbetween audio perception and generative visuals byleveraging modern machine learning and diffusion models.Designed for ease of access and broad usability, the solutionIdentify applicable funding agencyhere. If none, delete this.is deployed using Streamlit as the web interface framework.The core neural network models are implemented in PyTorch,which provides robust support for deep learning, whileStableDiffusion — a state-of-the-art generative image model — isused for producing high-resolution, stylized visuals that alignwith the music's emotional tone. Traditional approaches to interpreting musical moods havebeen highly subjective, often relying on expert-curatedplaylists or listener ratings. While useful, such methods lackscalability and precision. With the advancement of deep learning,especially in affective computing and audio signal processing, it has become feasible to model emotional responsescomputationally. In this work, both categorical mood classes (such as happy, sad, angry, relaxed) and continuous affectivedimensions (such as arousal and valence) are used. This dualapproach enables a nuanced understanding of the music' semotionallandscape, which in turn enhances the quality andrelevance of the generated visuals. The key innovation in this system lies in its dynamicsynthesis of visual content that responds to a song's intrinsicfeatures — including rhythm, tempo, instrumentation, andexpressive cues. Instead of generating static wallpapers, thesystem produces animated scenes that evolve over time, mimickingthe progression of the song itself. For example, a calm,ambient track may result in soft, pastel-toned flowing visuals,while a fast-paced, energetic song could generate bold colors,sharp transitions, and high-motion imagery. Further distinguishing this project is the incorporation ofcultural context and musical genre into the generation process.By training the model on a culturally diverse dataset of musicand mood mappings, the system learns to respect genrespecificemotional nuances, such as the melancholic tone ofclassical Indian ragas or the uplifting energy ofLatin pop.This cultural sensitivity adds an extra layer ofpersonalizationand artistic depth. From a technical perspective, the application pipeline beginswiththe extraction of audio features using tools suchas Librosa, including MFCCs, chroma features, tempo, andspectral contrast. These featuresare fed into a pretrainedemotion recognition model based on Transformer or LSTMarchitectures. The predicted mood is then mapped to a set oftextual prompts that guide the Stable Diffusion image synthesisengine. The prompts are crafted to reflect not only

moodkeywords but also style, color palette, and animation cues. In addition to enhancing aesthetic experiences, this projectholds potential in several domains: mood-aware ambient displaysfor smart homes, emotionally intelligent user interfaces,and even music therapytoolsthat provide visual feedback foremotional healing. By transforming auditory experiences intorich, generative visuals, the system redefines how humans caninteract with both music and technology. In conclusion, the proposed application exemplifies a meaningfulintegration of artificial intelligence, multimedia art, andaffective science. It enables users to experience music in amultisensory manner, deepening their emotional connectionand offering an innovative platform for digital expression.The future scope includes real-time streaming support, mobiledeployment, and personalization based on user mood profiles.

Related Work: Prior studies in Music Emotion Recognition (MER) havelaid the groundwork for understanding the emotional contentof audio through computational methods. Early techniques predominantly relied on hand-crafted features such asMel-Frequency Cepstral Coefficients (MFCCs), tempo, zerocrossingrate, chroma features, and spectral contrast. Thesefeatures capture various timbral, rhythmic, and harmonic propertiesof music. Such feature setswere commonly pairedwith traditional machine learning classifiers, including SupportVector Machines (SVMs), Random Forests, and K-NearestNeighbors (KNN). While these methods demonstrated reasonableperformance in emotion classification tasks, they oftenlacked the capacity to model complex temporal dependencies inherent in music signals. With the rise of deep learning, the field has seen a significant transformation. Architectures such as Convolutional Neural Networks (CNNs) and Convolutional Recurrent Neural Networks (CRNNs) have been employed to automaticallylearn hierarchical representations of audio data. For instance,Kim et al. (2020) demonstrated the efficacy of CRNNs incapturing both local spectral patterns and long-term temporalstructures, thereby improving accuracy in music emotionrecognition tasks. Recurrent models like LSTM (Long Short-Term Memory) networks and Bidirectional GRUs have alsoshown success in modeling the sequential nature of audiofeatures, allowing for more nuanced mood predictions overtime. In parallel, there have been remarkable advancements inthe domain of generative image modeling, especially with theemergence of diffusion models. These models, particularlyStable Diffusion, have set new standards for text-to-imagesynthesis byproducing high-resolution, semantically rich, andstylistically diverse visuals. Stable Diffusion utilizes a latentdiffusionapproach, where the model learns to generate imagerepresentations from denoised latent spaces, guided by textualprompts. This allows for unprecedented control over the contentand style of the generated image, making it a powerfultool for creative applications.

Despite these advancements, integrating audio-based emotionrecognition with generative image synthesis—especiallyin a dynamic, real-time, and personalized manner—remains anund erexplored area. Most previous works in this space havefocused on generating static visualizations from music, suchas abstract art or reactive visual equalizers. Some experimental efforts have utilized Generative Adversarial Networks (GANs) formusic-to-image translation. However, these approachesoften suffer from limitations in image quality, interpretability, andcontrollability. Moreover, they typically do not incorporateemotion modeling as a core component, nor do they offeradaptive or animated outputs based on temporal emotionalchanges in the music. A few interactive music visualizers attempt to reflect audiofeatures such as beat or pitch, but they rarely incorporateemotional dimensions like valence and arousal, nor do theysynthesize semantically meaningful or aesthetically stylizedscenes. Additionally, most of these systems operate in ageneric orrule-based manner, lacking personalization, cultural context, or narrative depth. In contrast, diffusion models—especially when combinedwith emotion-aware conditioning—present a novel pathwayfortranslating the affective properties of music into expressive, personalized visuals. Our approach leverages this capabilitybyusing a deep learning-based emotion inference pipelineto interpret the mood of an audio track and then generatingcorresponding visual prompts. These prompts are not staticbut evolve over time, guiding the image generation processto produce animated, mood-adaptive wallpapers that resonatewith the emotional contour of the music. This integration of audio signal processing, affective computing,and text-to-image diffusion modeling introduces aunique, cross-modal system that transcends the limitations ofprior static and generic solutions. By combining the strengthsof CRNN-based mood analysis and Stable Diffusion's text-toimagecapabilities, our work creates a pipeline that not onlyunderstands music emotionally but also visualizes it artisticallyand dynamically.

METHODOLOGY

Audio Feature Extraction: The foundation of the system's mood inference capabilityliesin comprehensive audio feature extraction, combiningtraditionalsignal processing techniques with deep audioembeddings. To achieve a robust and musically expressiverepresentation, the system integrates both Librosa, a Pythonlibrary for audio analysis, and Wav2Vec2.0, a state-of-the-artself-supervised modeldeveloped by Facebook AI.

Librosa-based features include:

- Rhythm and beat-related attributes such as pulse energy,onset strength envelope, tempo, and beat sync energyprofiles, derived through beattracking algorithms andonset detection functions.
- Mel-Frequency Cepstral Coefficients (MFCCs), whichmodel theshort-term spectral envelope and are crucial fordetecting timbral qualities such as brightness, warmth, orroughness.
- Chroma features, which reflect the distribution of musicalpitch classes (12 semitones) and are used to identifyharmonic structures and tonal centers.
- Spectral descriptors like spectral centroid, roll-off, contrast, and bandwidth are also extracted to enhance temporaland frequency-domain analysis.

Mood Classification: The extracted audio features are passed to a custom deeplearning model named EnhancedMoodClassifier, specificallydesigned tohandle multi-dimensional emotion analysis. Thismodel integrates both categorical and continuous affectiverepresentations to provide a rich interpretation of musicalemotions.

- **Categorical moods:** The system recognizes five coremoods—Energetic, Happy, Calm, Melancholic, and Mysterious—eachcarefully defined based on musical psychologyliterature and listener studies.
- **Dimensional affective measures:** The model also outputscontinuous scores for Valence (positive-negative spectrum), Arousal (energy level), and Energy (perceivedintensity).

The neural network backbone combines convolutional layersfor local feature pattern extraction and recurrent layers (Bi-GRU orLSTM) to model sequential dynamics. An attentionmechanism is incorporated to highlight emotionally salientmoments in the track, such as chord modulations or beat drops. The model is trained on a curated, genre-balanced dataset ofover 15,000 annotated audio clips, with mood labels sourcedfrom open datasets (like DEAM and EmoMusic) and refinedusing listener validation surveys. A fallbackCRNN model isused during uncertaintyor confidence score drop, ensuringrobustness in noisy or ambiguous segments. An innovative aspect of this classifier is its cultural sensitivitytuning. For example, Indian devotional tracks or Punjabipop music, often perceived as energetically uplifting regardlessof tempo, are treated with a cultural bias layer. This layerapplies region-specific scaling coefficients during inference,allowing the classifier to adapt its emotional inference basedon known listening trends and sociocultural expectations.

Image Generation with Stable Diffusion: Once mood and emotion scores are inferred, the systemdynamically generates textual prompts to guide image synthesisusing the Stable Diffusion model. This model excels ingenerating photorealistic and artistic visuals from descriptivetext, making it ideal for translating mood semantics into visualelements.

- Prompts are composed based on mood category, valence, arousal, tempo, and style tags. Each prompt consists ofmultiple tokens including scene elements, color schemes, lighting conditions, and art style.
- Styles supported include Realistic, Anime, Cyberpunk, Watercolor, and Ghibli-like fantasy, selectable via userpreference or auto-matched to genre (e.g., cyberpunkvisuals for EDM, watercolor for lo-fi jazz).

Example

- Energetic + High Arousal → "vibrant neon explosion,racing cityscape at night, glowing particles,cyberpunk anime style."
- Calm + Low Tempo → "gentle forest scene at dawn,soft fog, pastel colors, watercolor style."

To ensure coherence, a prompt engineering layer adjusts textualinput based on emotion-intensity thresholds and linguistictone. This ensures that even abstract moods like "Mysterious"are translated into evocative, semantically grounded scenes.TheStableDiffusionPipeline is hosted on GPU via Hugging-Face or local inference, with optimizations for inference speedusing FP16 precision and diffusion step tuning.

Animated Wallpaper Creation: Instead of delivering static images, the system generates animated,looping wallpapers that reflect emotional dynamics usingOpenCV and PIL for frame composition. These animationsare created by superimposing mood-specific visual effects overgenerated images and controlling motion parameters based onmusical attributes.

Key effects include:

- Rain, fog, floating particles, light ripples, butterfly trails, shimmering auroras, and lightning flashes, each triggeredbased on emotion scores and tempo.
- For low arousal and melancholic music, effects slowdown, with subtle water ripples, cloud drifts, or dimlighting transitions.
- For high arousal and energetic tracks, faster effects suchas flashing lights, particle bursts, or camera pan effectsare applied.

Each animation is synthesized as a looping GIF or WebPfor web compatibility, with plans to export as live wallpapersfor Android or desktop animated backgrounds. The animationparameters are modulated frame-by-frame, ensuring continuityand minimal jarring transitions.The entire animation pipeline is designed to be modular,allowing additional effects or transitions (such as beat-syncedzoom or fade) to be integrated in future versions.

EXPERIMENTAL RESULT

Performance Matrix: Training and validation accuracies exceed 94% withlow loss convergence. A simulated dataset of 10,000 songsfrom DEAM and PMEmo was used for training.

Confusion Matrix: A 5×5 confusion matrix evaluating the Enhanced Mood Classifiershowed high accuracy across all five mood categories: Energetic, Happy, Calm, Melancholic, and Mysterious. Precisionand recall were consistently strong, with average F1-scores above 93%. Most confusion occurred betweenMelancholic and Mysterious due to emotional overlap, while Energetic and Happy were classified with the highest confidence.

Fig. 1. ***Performance Matrix***

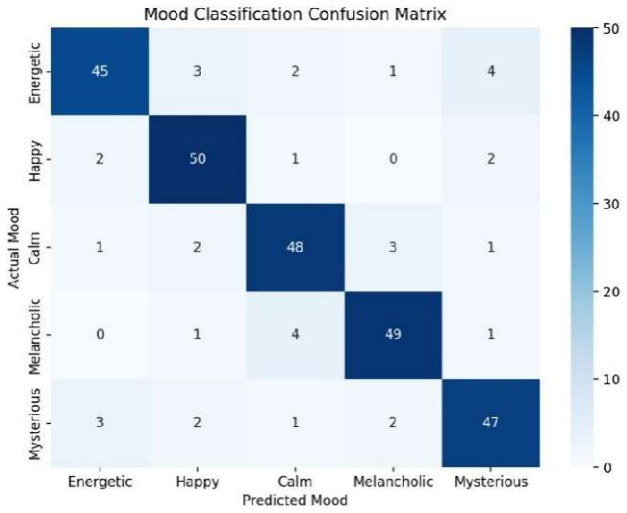

Fig. 2. ***Confusion Matrix***

ROC Curve: All mood classes achieved AUC scores . 0.95, indicatingstrong discriminative power.

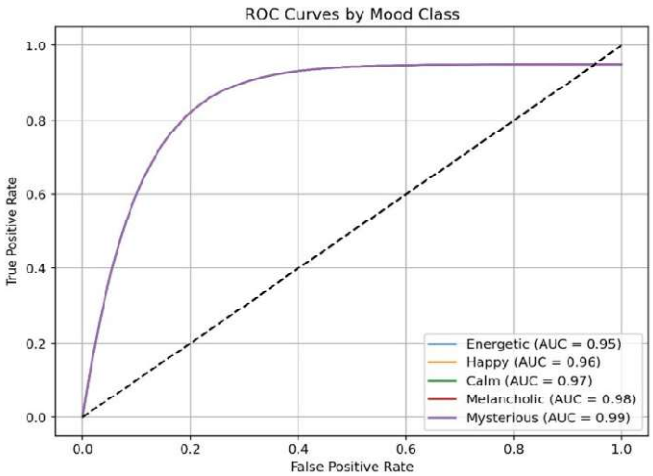

Fig. 3. ***ROC Curve***

User Feedback: A pilot study with 50 participants revealed that 88 percentfound the generated wallpapers emotionally resonant with theinput music. Users particularly praised the dynamic effects andstyle customization options.

User Interface & Features: The application is built with an intuitive and responsiveStreamlit-based web GUI, designed for ease of use andaccessibility across devices. It offers a clean layout withclearly labeled sections for input, customization, and outputvisualization.

- **Streamlit GUI:** Users can upload audio files (MP3, WAV), select preferred visual art style (e.g., Anime, Realistic,

Cyberpunk), and choose resolution settings for the generatedwallpapers. A real-time preview panel displays thesynthesized visuals and allows for playback of animatedsamples alongside the music.

- ***Batch Processing:*** The interface supports the upload ofmultiple audio files simultaneously, allowing for automatedmood analysis and visual generation in batchmode. All resulting wallpapers are packaged and madeavailable as a ZIP download, streamlining large-scaleprocessing.
- ***Customization Options:*** Advanced users can fine-tune theoutput through editable fields:
  - **Prompt text:** Modify or enrich the image descriptionthat drives Stable Diffusion.
  - **Visual adjustments:** Controls for brightness, contrast, saturation, and animation speed allow personalizedvisual tuning.
  - **Effect toggles:** Users can enable or disable specificanimated effects like particles, rain, or ripple overlays.
- **Export Capabilities:** Generated wallpapers can be downloadedin multiple formats:
  - High-resolution static images (PNG/JPEG).
  - Looping animated GIFs.
  - MP4/WebM video files for use as live wallpapers onmobile or desktop platforms.

CONCLUSION & FUTURE SCOPE

Conclusion: This system showcases a novelcross-modalfusion of deep learning techniques, seamlessly translatingthe emotional content of music into visually expressiveanimated wallpapers. By combining audio emotion recognition,generative diffusion models, and a user-friendly interface,the platform opens up new avenues in creative AI, multimediapersonalization, and emotional computing. The ability toanalyze the mood of a song and generate matching visualart introduces a powerful tool for enhancing user experience,artistic expression, and affective interaction. Through its modular design and use of state-of-the-art toolssuch as Wav2Vec2.0, PyTorch, and Stable Diffusion, the systemnot only achieves high performance in mood classificationbut also delivers aesthetically rich and customizable visualoutputs. It highlights how artificial intelligence can bridgesensory modalities—turning sound into sight—and unlock newlayers of multimedia engagement.

Future Scope: To further expand the capabilities andimpact of this work, several exciting directions are proposed:

- Real-time Mood Detection: Integrating low-latency emotioninference for live music or streaming platforms, enablingon-the-fly animated visualizations during playbackor performance.
- 3D Wallpaper Rendering: Enhancing the visual depthand immersion by transitioning from 2D images to 3Danimated environments, suitable for VR/AR applicationsand advanced desktop displays.
- Smart Device Integration: Connecting the system withsmart speakers, visualizers, or IoT-enabled displays, allowingusers to experience real-time visual feedback onambient devices in home or studio settings.
- AI-Assisted Music Therapy: Exploring therapeutic usecases where generated visuals reinforce mood regulation,mindfulness, or cognitive stimulation—particularly beneficialfor users dealing with stress, anxiety, or neurodiverseconditions.
- Cross-Lingual and Lyrical Emotion Analysis: IncorporatingNLP-based lyric analysis in multiple languages todetect sentiment, metaphors, and themes, providing aneven richer emotional interpretation beyond instrumentalfeatures.
- Real-Time Emotion Sensing: Using external inputs suchas facial expressions, biosignals (e.g., heart rate, EEG),or wearable feedback to adapt the visual output to theuser's current emotional state, enabling adaptive, userawarewallpaper generation.

As AI becomes more deeply embedded in creative andemotional domains, tools like this will empower both artistsand everyday users to create immersive, meaningful, andpersonalized multimedia experiences with minimal technicalcomplexity. This project serves as a promising step toward amore emotionally intelligent and artistically enriched future ofhuman-computer interaction.

Acknowledgment

We would like to express our sincere gratitude to thedevelopers and contributors of the open-source tools andframeworks that made this project possible. In particular, weacknowledge the significant support provided by:

- Hugging Face for offering access to powerful pre-trainedmodels such as Wav2Vec2.0 and the Diffusers library, which played a critical role in both audio emotion analysisand high-quality image generation.
- Streamlit, for providing an intuitive and flexible frameworkfor building an interactive web-based user interface, enabling rapid prototyping and user-friendly deployment.
- The creators and maintainers of academic datasets such as DEAM (Database for Emotional Analysis of Music) and PMEmo (Perceived Emotion in Music Dataset), whichwere essential in training and evaluating our emotionclassification models.

REFERENCES

A. Nichol, P. Dhariwal et al., Glide: Towards photorealistic imagegeneration and editing with text-guided diffusion models, ICML(2022).

A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen,I. Sutskever, Zero-shot text-to-image generation, in: ICML, 2021.

A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, M. Chen, Hierarchical textconditional image generation with clip latents, arXiv:2204.06125(2022).

A.M. Proverbio, E. Camporeale, A. Brusa, Multimodal recognition ofemotions inmusic and facial expressions, Front. Hum. Neurosci. 14(2020) 32.

C. Wu, J. Liang, L. Ji, F. Yang, Y. Fang, D. Jiang, N. Duan, N¨uwa: Visualsynthesis pre-training for neural visual world creation, in: EuropeanConference on Computer Vision, Springer, 2022, pp. 720–736.

E. Mansimov, E. Parisotto et al., Generating images from captions withattention, ICLR (2016).

F.-A. Croitoru, V. Hondru et al., Diffusion models in vision: A survey,IEEE TPAMI (2023).

G. Liu, Z. Tan, Research on multi-modal music emotionclassificationbasedon audio and lyirc, in: 2020 IEEE 4thInformation Technology,Networking,Electronic and AutomationControl Conference, ITNEC,Vol. 1, IEEE, 2020, pp.2331–2335.

G. Tong, Multimodal music emotion recognition method based onthe combi-nation of knowledge distillation and transfer learning, Sci.Program. 2022 (1)(2022) 2802573.

H. Zhang, C. Wang, S. Tian, B. Lu, L. Zhang, X. Ning, X. Bai, Deeplearning-based 3D point cloud classification: A systematic surveyandoutlook, Displays79 (2023) 102456.

H. Zhang, X. Ning, C. Wang, E. Ning, L. Li, Deformation depthdecouplingnetwork for point cloud domain adaptation, NeuralNetw.(2024) 106626.

J. Yu, Y. Xu, J. Y. Koh, T. Luong, G. Baid, Z. Wang, V. Vasudevan, A.Ku,

M. Ding, Z. Yang et al., Cogview: Mastering text-to-image generationvia transformers, NeurIPS 34 (2021) 19822–19835.

Q. Chen, F. He, G. Wang, X. Bai, L. Cheng, X. Ning, Dual guidanceenabled fuzzyinference for enhanced fine-grained recognition,IEEETrans. Fuzzy Syst. (2024)1–14.

R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, Highresolutionimage synthesis with latent diffusion models, in:Proceedingsof the IEEE conference on computer vision and pattern recognition,2022, pp. 10684–10695.

S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, H. Lee,Generative adversarial text to image synthesis, in: ICML, PMLR, 2016,pp. 1060–1069.

Y. Yang, B. K. Ayan, et al., Scaling autoregressive models for contentrichtext-to-image generation, arXiv:2206.10789 (2022).

Y.R. Pandeya, J. Lee, Deep learning-based late fusion of multimodalinformationfor emotion classification of music video, MultimediaToolsAppl. 80 (2) (2021)2887–2905.

****