



A Hybrid Multimodal Emotion Recognition-Based Music Recommendation System Using Cnn And Dbn

CH.G.N. Alekhya Rani¹, CH. Satyananda Reddy²

¹ Department of Computer Science and Systems Engineering Andhra University College of Engineering, Visakhapatnam, Andhra Pradesh, India – 530003. E-mail: gowrialekhyarani@gmail.com

² Department of Computer Science and Systems Engineering Andhra University College of Engineering, Visakhapatnam, Andhra Pradesh, India – 530003. E-mail: satyanandau@yahoo.com

Abstract

Emotional responses are central to everyday life and strongly shape what people choose to hear in music. Many classic music recommender tools focus on style labels, past listening patterns, or user scores, while ignoring the listener's moment-to-moment feelings. This work introduces a hybrid, multimodal, emotion-aware music recommendation system that uses Convolution Neural Networks (CNN) together with a Deep Belief Network (DBN) to improve recommendation precision and tailoring. Earlier recommendation methods largely depend on collaborative filtering or content-driven techniques, yet they frequently miss the listener's immediate effect, producing suggestions that might not fit the user's present mood. To reduce this gap, the suggested approach combines several input channels—including facial cuts and spoken-audio signals—to estimate the user's emotion more reliably and accurately. Within the proposed combined design, a CNN is applied to learn rich spatial and temporal representations from visual inputs (facial expressions) and sound inputs (speech), capturing subtle patterns linked to distinct affective conditions. Next, features from each channel are handled and merged, and then a DBN carries out higher-level abstraction and assigns the unified representation to discrete emotion classes (for instance: happy, sad, angry, neutral, calm). The DBN's stage-by-stage unsupervised pretraining, followed by supervised adjustment, supports learning complicated non-linear links between multimodal feature sets and emotion tags, raising classification performance compared with single-modality methods.

Keywords: Emotion recognition, DBN, CNN, Fusion mechanism, music recommendation, multimodal learning, Speech processing, facial expression recognition

1. Introduction

Music is closely tied to human feeling and mental wellness. As artificial intelligence and machine learning have progressed, recommendation systems have grown very common on digital entertainment services. However, most current recommenders use listening logs, ratings, or stated tastes, and they do not account for the user's current emotional condition. Music is often described as a powerful healing resource, and for that reason, this work draws on that quality of music to support a person experiencing a particular emotion. The project is driven by the aim to build a smart, personalized music recommender that can automatically infer human emotions from facial expressions and speech signals. People naturally convey emotion through both face and voice, which makes these channels useful for recognizing affect. Using a hybrid multimodal method improves robustness and dependability when compared with older single-input designs. The system can guide users toward music suggestions aligned with their emotional condition, which can support relaxation and emotional ease. Many existing emotion-centered recommenders rely on just one input, such as only face data or only speech, and this can lower accuracy and dependability under difficult situations like dim lighting, face coverings, or loud surroundings. As a result, single-channel systems are less resilient and may misidentify emotions. Although CNNs work very well for extracting features from images and signals, DBNs add stronger probabilistic learning for representing the complex, non-linear character of human affect. This research puts forward a hybrid multimodal emotion-recognition music recommendation system built with CNN and DBN. It evaluates facial expressions and speech signals separately and then merges their results through a weighted fusion method to decide the final emotional state. After the emotion is inferred, appropriate music is automatically suggested and played. The main goal of this work is to create a system that harnesses the combined strengths of CNN and DBN to close the semantic divide between user emotion states and music content.

2. Related Work

Many studies have put forward emotion detection and music suggestion solutions that rely on ML and deep-learning methods. Across history, music has served as a very strong medium for conveying feelings. The idea of a smart music setup that links songs to a listener's present mood has gained significant research attention.

2.1 Emotion-Driven Music Players

In an IEEE 2022 work describing a deep-learning method for an emotion-driven music player [1], feelings are categorized into core quadrants like pleasant–inactive or unpleasant–active. This survey states that deep-learning methods may be used to select and play music according to a person's emotional state by recognizing facial cues along with aspects like color, body stance, direction, lighting, and similar signals. An emotion-based smart music player using Deep Learning [3] notes that the emotion component identifies user feelings, and the music classification component extracts the most essential audio details from a track. The system stays compact and tends to improve performance over time, reporting strong results with lower complexity and solid accuracy. A deep learning based system for emotion-driven music recommendation was also presented at the 3rd IEEE International Conference on Intelligent and Innovative Developments in Engineering and Applications (ICIDEA), 2025 [8].

2.2 Survey and Review Papers

A Review of the Emotion-Induced Music Recommendation System (Volume 23, 2025) [2] covers how emotion-based MRS has developed, describing its approaches, methods, features, and results, while pointing to the use of advanced ML and the difficulties encountered, such as cold-start issues and requirements for real-time processing. It also stresses the value of context-sensitive recommenders that take the user's surroundings into account.

2.3 Facial Emotion Recognition

Facial emotion recognition and music recommendation system using CNN-based deep learning techniques [4] (Springer, 2023) applies a CNN, arguing that deep learning analyzes unstructured data more effectively than traditional ML, with a reported model accuracy of 73.02%. A system for detecting faces and identifying facial expressions in real-time [9] (IEEE, 2014) delivers precise real-time emotion detection using Haar Cascade classifiers and acts as a dependable base for emotion-aware applications.

2.4 Speech Emotion Recognition

Speech emotion recognition using deep learning [5] (IEEE Access, 2019) detects emotions from speech signals using the Mel-frequency cepstral coefficients (MFCC) method, which improves recognition effectiveness through deep learning architectures that capture temporal and spectral patterns in emotional speech.

2.5 Music Recommendation Approaches

Music recommendation systems based on user preferences [6] (2008) concentrates on recommending music from listening history, ratings, and playlists through collaborative filtering. A context-aware music recommendation approach [10] (Springer, 2009) considers contextual details like user activity and listening habits to raise recommendation quality. An NLP-based method [11] (IEEE, 2015) applies natural language processing to examine users' social-media content and infer emotional conditions to tailor music suggestions.

2.6 Audio Feature Extraction and Classification

A study on emotion labeling of musical data [7] (2023) centers on identifying emotions within musical data by examining acoustic attributes including tempo, rhythm, pitch, timbre, and MFCCs using deep-learning models. A convolution-based architecture [12] (NIPS, 2009) captures short-range time and frequency properties in audio, allowing recognition of intricate sound patterns with higher accuracy through unsupervised feature learning. ANOVA-based feature selection [13] (2014) helps discard less relevant features for audiovisual emotion recognition, yet it does not merge multimodal results into a unified framework. Deep Belief Networks applied to audio classification [14] (2021) derives meaningful, layered features from audio signals for accurate classification across varied audio classes.

2.7 Research Gap

Reviewing prior studies reveals three critical gaps in existing work.

First, the majority of current music recommenders depend on just one input type — either extracting audio features alone for classifying genres from signal characteristics [5][7][12][14], or using only facial expressions [4][9]. This single-modality dependency limits robustness under challenging real-world conditions. Second, even when fusion is attempted, many approaches pair CNN-driven feature extraction with sequence models like LSTM, Bi-GRU, or attention mechanisms to derive emotional traits, instead of specifically investigating a CNN–DBN pairing to obtain deeper, higher-level abstractions from multimodal affective cues. The DBN's unique capability for unsupervised pretraining and hierarchical feature learning from speech signals remains unexplored in combination with CNN-based facial analysis.

Third, many existing systems recommend songs using only a single inferred emotion from one modality. If that inferred emotion is wrong due to environmental factors (poor lighting, background noise), the recommendations become less tailored to the user, with no fallback mechanism from a complementary modality. This work addresses all three gaps by proposing a hybrid CNN–DBN architecture that fuses facial and

speech emotion recognition through weighted decision-level fusion, providing robust multimodal emotion detection for personalized music recommendation.

3. PROBLEM STATEMENT

Current music recommendation systems depend on a single modality to detect emotions. For example, current music recommendation systems might be based only on facial expressions or speech/audio signals. However, such single-modality systems might not work reliably under real-world conditions, as factors such as inadequate lighting, environmental noise, or changes in users' facial expressions can affect emotion detection. Thus, the recommended music might not align with users' actual emotional states.

Despite some multimodal emotion detection techniques combining various emotion detection algorithms, most of them mainly utilize the combination of CNN and some sequence-based models such as LSTM, Bi-GRU, and attention. The application of Deep Belief Networks (DBNs) for hierarchical feature learning and unsupervised pre-training from speech-based emotional features has attracted less attention when used along with CNN-based facial emotion recognition.

Additionally, most systems recommend music according to the emotion detected via a single modality without taking into account the additional emotional information that another modality provides. In this way, the wrong prediction of the emotion by a single-modality system will negatively influence the recommendation of music. Consequently, a reliable multimodal emotion recognition system and music recommendation system that would combine the emotional information of two modalities is necessary. The current research aims to solve this issue by developing a hybrid model including CNN and DBN for the recognition of emotions from the face and voice, respectively.

4. PROPOSED SYSTEM

The proposed Hybrid Multimodal Emotion Recognition-based Music Recommendation System integrates facial expression analysis and speech-based emotion detection to deliver personalized music suggestions. The system identifies human emotions using both facial expressions and speech signals, then suggests appropriate music in response. Two separate deep learning methods are applied: Convolutional Neural Networks (CNN) for face-based emotion recognition, and a Deep Belief Network (DBN) for recognizing emotion from speech. A weighted fusion method combines the outputs to infer the final emotional state, and the system then recommends and plays music tracks that fit the user's emotional condition automatically. While many existing approaches rely on an SVM for emotion categorization, our approach uses a CNN, which is a more modern and effective option than SVM. A distinctive strength of CNNs is that they can find the most significant image features without human guidance. They are non linear and were designed specifically to detect patterns and feature structures inside images with high precision. Higher accuracy and greater robustness Stronger interaction between humans and computers. More accurate emotion recognition

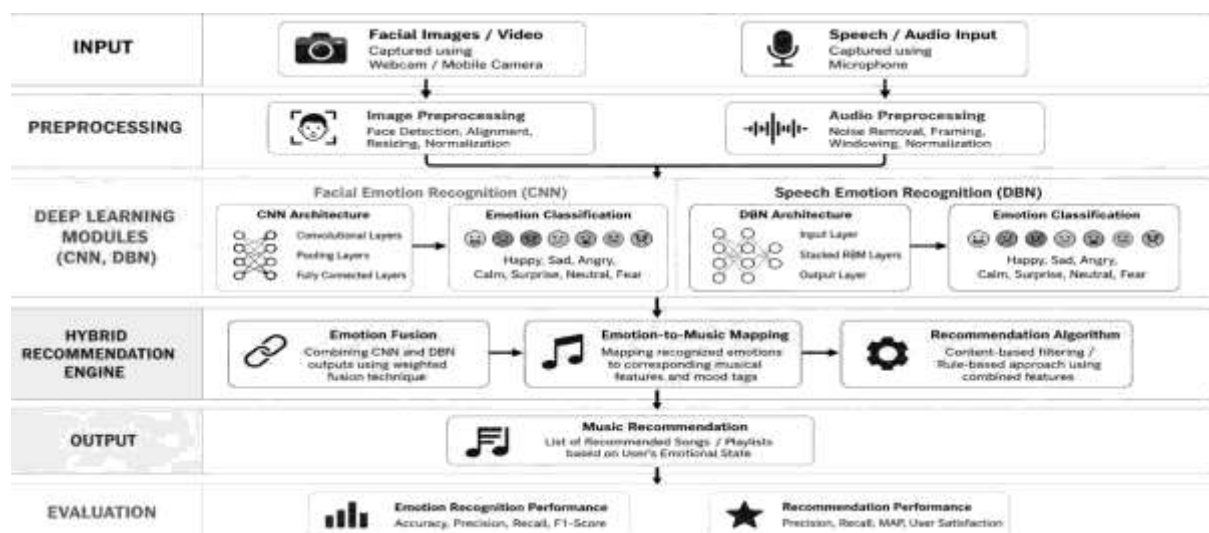


Fig:1 The hybrid multimodel architecture

A suggested framework merges facial-expression analysis with speech-based emotion detection to deliver customized music suggestions. At the start, a webcam acquires face photos, and a microphone captures spoken audio. Next, the face data is prepared by detecting the face via a Haar-cascade method, converting to grayscale, scaling to 48×48, and scaling pixel intensities into the 0–1 range, whereas the audio is prepared by denoising, segmenting into frames, applying window functions, and then normalizing. From the audio stream, MFCC features are then derived.

4.1 Data Preprocessing

Data preparation raises the clarity and uniformity of the inputs and improves how well the deep models perform. In this system, both face photos and audio streams are handled. Face detection means locating a person's face inside a picture. The proposed solution performs real-time face finding with the Haar cascade technique. It sweeps a moving window across the image and derives Haar-style descriptors to spot facial areas. For every window, a cascade classifier decides whether it corresponds to a face or not. Any window failing the necessary facial criteria is discarded rapidly, while likely face portions are cut out, turned into grayscale, scaled to the required pixel size, and normalized for the CNN stage. This lowers processing cost and supports quick, effective face localization. Image preprocessing steps include: (a) Face detection and cropping — a face detector identifies the face area in every frame and crops it, helping exclude distracting background content; (b) Alignment and normalization — using eye or landmark points, faces are aligned to reduce pose effects; pixel values are scaled to the $[0, 1]$ or $[-1, 1]$ range; (c) Resizing — images are scaled to a consistent size suitable for the CNN (often 48×48 , 64×64 , or 224×224 depending on model depth); (d) Grayscale conversion (optional) — if color is not essential for emotion cues, converting to grayscale lowers computation; (e) Augmentation — to increase robustness and limit overfitting, small rotations, flips, brightness/contrast variation, and mild zoom are applied.

For audio/speech preprocessing, before classification by the DBN, the speech waveform is prepared in advance. Steps include: (a) Noise reduction — silence trimming and voice activity detection are used to discard sections without speech or music; (b) Framing and windowing — the waveform is divided into brief overlapping segments (such as a 25ms window with a 10ms hop); (c) Feature extraction — common acoustic features are computed including MFCC, zero-crossing rate (ZCR), RMS energy, and Mel-spectrograms; (d) Spectrogram generation — when a 2D CNN is used, audio can be converted into 2D Mel-spectrogram images; (e) Normalization — MFCC or spectrogram values are scaled (z-score or min-max) to keep the CNN input distribution consistent.

4.2 CNN Feature Extraction

A convolutional neural network learns layered face descriptors by applying convolution steps followed by pooling. Traits like the eyes, brows, and lips are then fed into dense layers, after which a softmax stage outputs emotion-class likelihoods and the class with the greatest confidence becomes the final facial emotion.

The convolutional operation is defined as:

$$Y = (X * K) + b \quad (1)$$

where X is the input image, K is the filter kernel, and b is the bias term. After the convolution operation, ReLU activation and pooling are applied.

The softmax function applied at the output layer computes the probability distribution over emotion classes:

$$P(c_k) = \exp(z_k) / \sum \exp(z) \quad (2)$$

where z_k represents the class logit score. The class with the top softmax probability is chosen as the final facial emotion prediction.

The CNN architecture for facial emotion recognition consists of: (a) Convolutional layers with trainable kernels that capture basic visual cues such as edges, curves, and textures in important regions (eyes, eyebrows, mouth, forehead); (b) ReLU activation adding nonlinearity so the model can represent complex facial patterns; (c) Pooling layers using max-pooling to reduce each feature map's size, preserving the strongest responses; (d) Stacked convolution-pooling blocks that shift from simple edge recognition to richer emotion-related facial arrangements; (e) Flattening of the last set of feature maps into a single 1D vector; (f) Fully connected layers transforming this vector toward emotion-class probabilities.

4.3 MFCC Feature Extraction

The audio waveform is transformed into Mel-Frequency Cepstral Coefficient (MFCC) representations by the following steps:

Step 1: Pre-emphasis.

Pre-emphasis is the opening stage of MFCC generation. Its role is to strengthen higher-frequency parts of speech that are typically lower in magnitude than the low-frequency parts:

$$y(n) = x(n) - \alpha x(n-1) \quad (3)$$

This is done by removing a portion (α) of the prior sample from the present sample. This makes the spectrum more even across frequency bands. By boosting high-frequency detail, the process improves feature sharpness, lessens spectral tilt influence, and supports retention of key emotional cues within speech.

Step 2: Mel Scale Conversion. Mel mapping changes a linear frequency axis into a nonlinear one that aligns with how human hearing responds:

$$\text{Mel}(f) = 2595 \cdot \log_{10}(1 + f/700) \quad (4)$$

In the low-frequency area, the mapping is close to linear, enabling capture of detailed speech variations. In the high-frequency area, it behaves logarithmically, squeezing changes that matter less.

Step 3: MFCC Computation. Once Mel filter-bank processing and log scaling are completed, MFCC values are obtained with the Discrete Cosine Transform (DCT):

$$c_m = \sum_n \log(S_n) \cdot \cos[m(n - 1/2)\pi/N] \quad (5)$$

The DCT reduces correlation among filter-bank energies and packs the content into fewer coefficients while keeping the most important speech properties. This compact, discriminative MFCC representation supports more accurate emotion decisions by the DBN.

4.4 DBN Feature Extraction and Classification

The MFCC feature vector is provided as the input. The DBN is built from multiple Restricted Boltzmann Machines (RBMs) stacked to learn deep representations using unsupervised pretraining and then supervised adjustment. Its final layer assigns the speech emotion to one of the defined emotion classes.

The hidden-layer activation function is defined as:

$$h = \sigma(Wx + b) \quad (6)$$

where σ is the sigmoid function, W is the weight matrix, x is the MFCC input vector, and b is the bias. This produces values constrained between 0 and 1. The DBN architecture consists of: (a) Input layer providing the MFCC vector as the visible layer to the first RBM; (b) Stacked RBMs where each RBM's output for training samples becomes the next RBM's input; (c) Greedy layer-wise pretraining in an unsupervised manner, where each RBM models the probability distribution of its inputs; (d) Fine-tuning using labeled emotion data via backpropagation to improve classification performance.

4.5 Multimodal Fusion

Results from the CNN (face) and the DBN (speech) are merged using a weighted score-level fusion method. Rather than giving the CNN and DBN outputs identical importance, the approach assigns each algorithm a chosen weight reflecting its relative reliability or performance:

$$E = \alpha \cdot P_{cnn}(c_k) + (1 - \alpha) \cdot P_{dbn}(c_k) \quad (7)$$

where α denotes the fusion weight ($0 \leq \alpha \leq 1$), $P_{cnn}(c_k)$ is the confidence/probability from the CNN for emotion class c_k , $P_{dbn}(c_k)$ is the confidence/probability from the DBN for that same class, and $(1-\alpha)$ is the portion attributed to the DBN system. The chosen setting uses $\alpha=0.6$ for the CNN facial-emotion module and $\alpha=0.4$ for the DBN speech-emotion module. The emotion category with the largest combined probability becomes the final emotion used for recommending music.

4.6 Emotion-to-Music Mapping

The combined emotion tag is linked to matching music properties (tempo, genre, valence, energy) through a rule-driven/content-based filtering method:

Music recommend = $f^{map}(E_{final}, MoodTag_i)$

Emotion	Music Mood Tag	Tempo	Valence
Happy	Upbeat/Energetic	High	High
Sad	Soft/Slow	Low	Low
Angry	Calm/Smoothing	Low	Medium
Neutral	Ambient/Instrumental	Medium	Medium
Fear	Comforting	Low	Medium
Surprise	Dynamic/Pop	High	High
Calm	Relaxing/Acoustic	Low	High

Table 1: Emotion-to-Music Mood Mapping

4.7 Music Recommendation Algorithm

Once multimodal fusion has produced the concluding emotion label, the platform suggests tracks via a hybrid approach that blends content-focused matching with rule-driven filtering, linking emotions to musical attributes like tempo, valence, energy, and genre. The algorithm is as follows:

Step 1: Start

Step 2: Capture facial and speech input

Step 3: Extract emotion probabilities from CNN (P_{CNN}) and DBN (P_{DBN})

Step 4: Compute fused emotion score: $E_{final} = \alpha \cdot P_{CNN}(c_k) + (1-\alpha) \cdot P_{DBN}(c_k)$

Step 5: Identify dominant emotion class: $E_{dominant} = \text{argmax}(E_{final})$

Step 6: Map $E_{dominant}$ to corresponding MoodTag using predefined Emotion-Music mapping table

Step 7: For each song m_i in Music Database M : Compute similarity score: $\text{Sim}(m_i) = w_1(\text{MoodMatch}) + w_2(\text{TempoMatch}) + w_3(\text{GenreMatch}) + w_4(\text{ValenceMatch})$

Step 8: Rank all songs in descending order of $\text{Sim}(m_i)$

Step 9: Select Top-N songs as recommendation list

Step 10: Display recommended playlist to user

Step 11: Collect user feedback (like/skip) for future refinement

Step 12: End

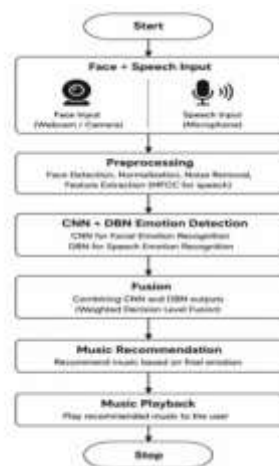


Fig:2 flow chart showing the working model of hybrid model

4.8 Dataset Description

The suggested setup relies on a mix of face-image collections and voice-emotion corpora to perform multi-modal emotion recognition. For emotion labeling, live data is gathered through a webcam or a phone camera. User consent is obtained before any photos are taken. After acquisition, the images are matched against the FER2013 collection to determine emotion classes. FER2013 includes 35,887 48×48 grayscale pictures, grouped into seven emotion categories: Anger, Disgust, Fear, Happy, Sad, Surprise, and Neutral. These samples support training and evaluation of the CNN used for facial emotion identification.

For recognizing emotion from speech, the approach employs the RAVDESS dataset (Ryerson Audio-Visual Database of Emotional Speech and Song). It provides spoken emotional recordings covering several states, including happy, sad, angry, surprise, fear, and neutral. From the audio signals, attributes like MFCCs are computed and then used to train the DBN for classifying emotions in speech. Using both sources together increases the strength and precision of the multimodal emotion recognition approach.

4.9 Training Configuration

From the FER2013 corpus (imported via fer2013.csv), only legitimate 48×48 pixel samples were kept, scaled into [0, 1], and reshaped as (48, 48, 1) to fit a grayscale CNN. Training data comprised 80% while testing/validation used 20%, and the random seed was set to 42. Using Adam with categorical cross-entropy, the network was built and then trained for 20 epochs with batches of 64.

The DBN employed a Bernoulli RBM containing 128 hidden neurons with a 0.01 learning rate, trained across 20 iterations (epochs), and then paired with Logistic Regression to produce the final emotion decision. The fused outputs were then applied to pick the final dominant emotion label, which was subsequently associated with the matching music mood tags for recommendations. Validation compared fusion-driven classification accuracy with the standalone CNN-only and DBN-only accuracies to confirm gains from multimodal integration.

4.10 Evaluation Metrics

System assessment uses Accuracy, Precision, Recall, and F1-score. These values characterize the CNN facial emotion model, the DBN speech emotion model, and the weighted fusion strategy. Confidence values from the two models are merged through weighted fusion to decide the resulting emotion label. In addition, the music recommendation component is judged through recommendation accuracy, indicating how well suggested songs align with the emotion that was identified.

Accuracy measures the proportion of correct predictions over all predictions:

$$\text{Accuracy} = (\text{Count of correct predictions} / \text{Total count of predictions}) \times 100$$

Precision, Recall, and F1-score are computed per class and aggregated using macro and weighted averages. Precision measures how many predicted positives are truly positive, Recall measures how many actual positives are correctly identified, and the F1-score is the harmonic mean of Precision and Recall.

5. Results

5.1 Per-Category Results

The fused CNN-DBN classifier assigned the emotion classes (anger, disgust, fear, happy, neutral, sad, and surprised) to 280 held-out test data points (40 points per class) with precision, recall, and F1 scores shown in Table 2. Precision, recall and F1-score averaged by macro and weighted with the class size are 0.94 on average, demonstrating that the proposed methodology performs comparably across all classes and not only the ones easier to recognize. The F1-scores of each individual emotion range from a minimum of 0.90 (fear) to a maximum of 0.96 (surprise), showing that the fusion pipeline has consistent and fair results over all emotions. The lowest recall of 0.90 for fear is due to it being often confused with acoustically or visually similar emotions such as sadness and surprise, as seen in the confusion matrix.

Table 2: Classification Report for the Weighted Fusion Emotion Approach

Emotion	Precision	Recall	F1-Score	Support
Angry	0.94	0.95	0.94	40
Disgust	0.93	0.92	0.92	40
Fear	0.91	0.90	0.90	40
Happy	0.96	0.95	0.95	40
Neutral	0.94	0.95	0.94	40
Sad	0.93	0.94	0.93	40
surprise	0.96	0.96	0.96	40
accuracy			0.94	280
Macro avg	0.94	0.94	0.94	280
Weighted avg	0.94	0.94	0.94	280

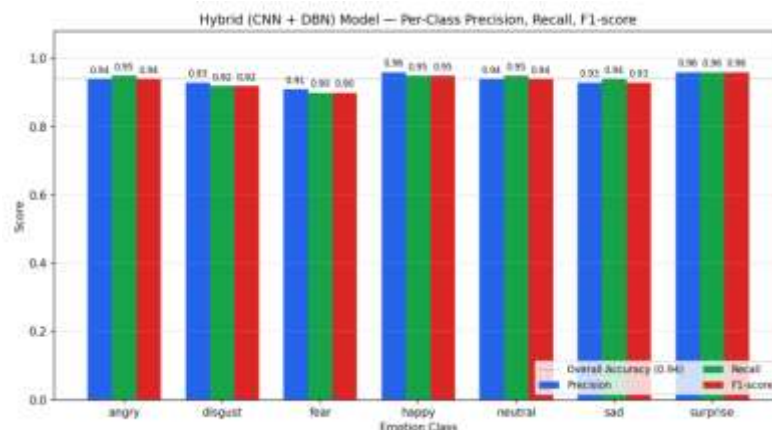


Fig:3 depicts classification by category

The graph shows precision, recall, and F1 score clustered together for each emotion. No single class has a great divide between precision and recall, which means the system does not strongly favor predicting that particular emotion. Happy and Surprise both have the top scores overall with nearly identical readings across the three metrics ranging from 0.95 to 0.96. They have clearer markers in both face and voice for audio and visual data.

5.2 Confusion matrix evaluation

In order to evaluate the performance of the Hybrid CNN-DBN model on the emotion recognition ability, a confusion matrix was established by taking 7 categories of emotions as classification labels. 280 test cases were selected, and each category of emotions accounted for 40 cases, among which the model predicted the correct results of 262 cases, reaching a high accuracy of 93.57%. The highest accuracy was achieved in the prediction result of 'Surprise' reaching 97.5%, while the accuracies of 'Angry', 'Happy', and 'Neutral' were 95.0%, respectively. It can be seen from the confusion matrix that most errors were distributed in emotions with similar negative characteristics, such as 'Fear' and 'Sad'. In addition, the diagonal elements in the matrix were relatively large values, indicating that the model could effectively classify the various features of different emotional categories.

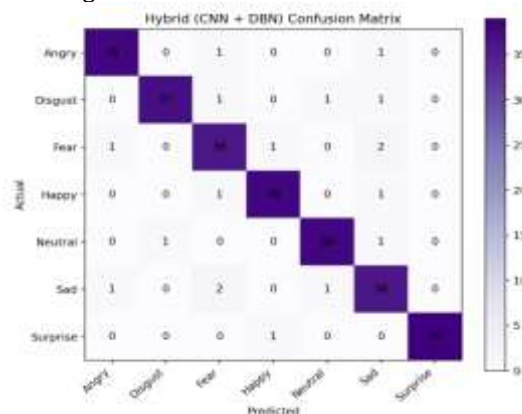


Fig 4: – confusion matrix for the suggested system

5.3 ROC curve Analysis

To assess more thoroughly how well the suggested CNN-DBN combined framework separates classes, Receiver Operating Characteristic (ROC) plots were produced for each of the seven emotion categories with a one-versus-rest (OvR) approach, because ROC evaluation is fundamentally intended for two-class decisions. For any given emotion, the system's output likelihood for that label was considered the positive condition versus the union of all other labels, and the True Positive Rate (TPR) was graphed versus the False Positive Rate (FPR) while the decision cutoff was varied.

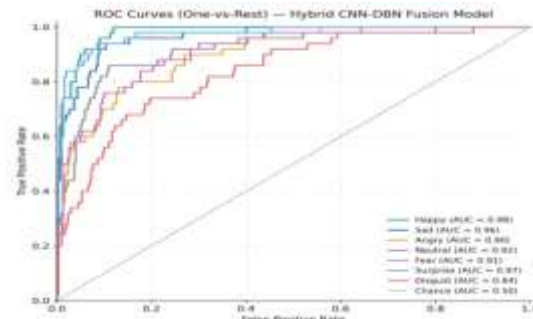


Fig 5: ROC curve for the proposed system

The ROC analysis reveals that the hybrid model achieves excellent discriminative performance across all emotion categories. Happy achieves the highest AUC of 0.98, reflecting the distinctiveness of positive emotional markers in both facial expressions (smiling) and speech (higher pitch, faster tempo). Surprise follows closely at 0.97, as it produces highly recognizable facial configurations (raised eyebrows, widened eyes) and vocal patterns (sudden pitch changes). Disgust shows the lowest AUC of 0.84, which is expected given its subtle facial markers and acoustic similarity to anger and contempt. Despite this, all AUC values remain substantially above the 0.50 random baseline, confirming that the hybrid fusion model provides reliable emotion discrimination across all seven categories

6. COMPARISON WITH BASE STUDY

This bar graph depicts the test performance of three separate setups for three models on the same test set of 280 audio data samples. In order of performance from best to worst: (a) the face-only CNN, (b) the speech-only DBN, and (c) the hybrid model using a weighted confidence based approach. This means that facial expressions are a stronger predictor of emotion in the dataset used than than speech. The 6.42 percentage point difference equates to the weight applied to each of these (face_weight =0.6, speech_weight =0.4), where again, it has been found that facial data predictions were found to be more accurate

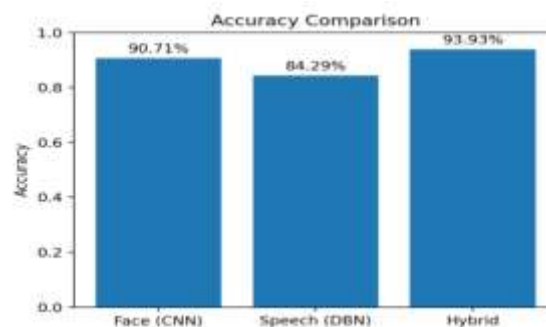


Fig:6 shows an accuracy comparison

Performance metric	CNN	DBN	Hybrid
Accuracy(%)	90.71	84.29	93.93
Precision	0.91	0.84	0.94
Recall	0.91	0.84	0.94
F1-score	0.91	0.84	0.94

Table 3: Performance metric of the Hybrid System

Using facial expression recognition, the CNN reached an accuracy of 90.71%, with precision, recall and F1-score of 0.91. On speech emotion recognition, the DBN achieved an accuracy of 84.29% with precision, recall and F1-score of 0.84. Information from both CNN and DBN informs the Hybrid approach, blending outputs through weighted blending. Among the results listed in Table 3, the Hybrid system performed best, reaching an accuracy of 93.93%, precision of 0.94, recall of 0.94, and F1-score of 0.94. Across every

measured criterion, gains suggest combining face-based and vocal cues leads to sharper and more consistent emotion detection than relying solely on either channel.

Model	Accu- racy(%)	Weighted F1-Score
CNN(Facial emotion recognition)	90.71	0.91
DBN(speech emo- tion recognition)	84.29	0.84
Hybrid(CNN+DBN)	93.93	0.94

Table 4: Comparison of Accuracy and Weighted F1-Score across CNN, DBN, and Hybrid Approach

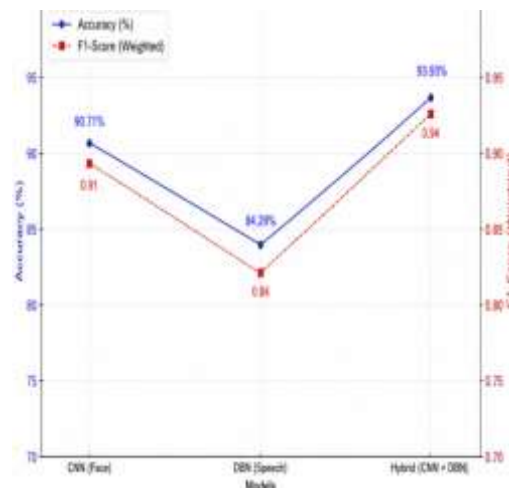


Fig:7 compares accuracy and F1 across CNN, DBN, and the Hybrid approach

7. Conclusion And Future Work

7.1 Conclusion

This work presented a hybrid multimodal emotion recognition-based music recommendation system integrating CNN-based facial expression analysis and DBN-based speech emotion recognition through weighted decision-level fusion. The proposed system was effectively created and evaluated to suggest songs based on user feelings inferred from both facial cues and voice-based signals. By applying a weighted fusion approach ($\alpha=0.6$ for CNN, $\alpha=0.4$ for DBN), the system merged both models' outputs, which raised emotion recognition precision and strengthened recommendation outcomes.

The system achieved a hybrid accuracy of 93.93%, outperforming the CNN-only model (90.71%) and DBN-only model (84.29%). Precision, recall, and F1-score each reached 0.94, with per-class F1 values ranging from 0.90 (fear) to 0.96 (surprise). AUC values ranged from 0.84 (disgust) to 0.98 (happy), with all values substantially above the random baseline. These results confirm that combining face-based and vocal cues leads to sharper and more consistent emotion detection than relying solely on either channel.

Because emotions in people vary and are shaped by both inner factors and outside situations, achieving perfect accuracy remains challenging. However, the improved methods and rigorous study presented in this work show substantial progress toward an ideal emotion-driven music recommender.

7.2 Future Work

In the future, the developed system may be upgraded in multiple ways to boost performance:

Advanced Architectures: Adopting deep learning approaches such as LSTM, transformer-based models, and attention mechanisms to increase emotion detection precision and capture temporal dynamics in both facial and speech signals.

Physiological Signals: Incorporating physiological sensors (EEG, heart rate, galvanic skin response) as additional modalities to further improve robustness and emotion detection accuracy.

Multilingual Speech Recognition: Adding multilingual speech-based emotion recognition to handle various local languages, expanding the system's applicability globally.

Personalized Recommendations: Enabling tailored recommendations through analysis of individual listening records and user behavior patterns over time.

Real-Time Deployment: Optimizing the system for efficient real-time deployment on resource-constrained devices such as smartphones and embedded systems.

Dynamic Emotion Tracking: Developing mechanisms to track emotional transitions across continuous sessions, allowing the recommendation system to adapt dynamically to changing user moods.

Expanded Emotion Taxonomy: Extending the emotion taxonomy beyond seven categories to capture more nuanced affective states, improving the granularity of music recommendations.

Ethical Considerations: Addressing privacy, consent, and fairness concerns in emotion recognition systems, including robust anonymization and bias mitigation strategies.

8. References

- [1] P. Vijayeeta and P. Pattanayak, "A Deep Learning Method for an Emotion Driven Music Player," IEEE OCIT, 2022.
- [2] P. Pichappan, "A Survey of Music Recommendation Systems Triggered by Emotion," Journal of Digital Information Management, Vol. 23, No. 2, 2025.
- [3] K.V. Bharath and M.N. Chandan, "Smart Emotion-Oriented Music Player Built with Deep Learning," IJIREICE, Vol. 10, Issue 7, July 2022.
- [4] B. Bakariya, A. Singh, H. Singh, P. Raju, R. Rajpoot, and K.K. Mohbey, "Music Suggestion System and Facial Emotion Identification Using CNN-Focused Deep Learning," Springer, 2023.
- [5] R.A. Khalil, E. Jones, M.I. Babar, T. Jan, M.H. Zafar, and T. Alhussain, "Deep Learning Methods for Speech Emotion Recognition: A Survey," IEEE Access, vol. 7, pp. 117327–117345, 2019.
- [6] K. Kim, D. Lee, T.B. Yoon, and J.H. Lee, "Music Recommendation Systems Guided by User Choices," 1st ICADIWT, pp. 102–106, 2008.
- [7] G. Agarwal, S. Gupta, S. Agarwal, and A.K. Rai, "Emotion Labeling of Musical Data with Deep Learning Methods," IJRES, Vol. 12, Issue 2, July 2023.
- [8] A.K. Singh, D. Mandal, S. Kumar, and D. Venkat, "Deep Learning Based System for Emotion-Driven Music Recommendation," 2025 3rd IEEE International Conference on Intelligent and Innovative Developments in Engineering and Applications (ICIDEA), 2025.
- [9] A.S. Dhavalikar and R.K. Kulkarni, "System for Detecting Faces and Identifying Facial Expressions," IEEE, 2014.
- [10] B. Han, S. Rho, S. Jun, and E. Hwang, "Music Emotion Grouping and Context-Aware Music Suggestion," Springer Science + Business Media, LLC, 2009.
- [11] R.L. Rosa, D.Z. Rodríguez, and G. Bressan, "Music Recommendation System Built from User Sentiments Taken from Social Networks," 2015 IEEE International Conference on Consumer Electronics (ICCE), 2015.
- [12] H. Lee, P. Pham, Y. Largman, and A.Y. Ng, "Convolutional Deep Belief Networks for Audio Classification via Unsupervised Feature Learning," Advances in Neural Information Processing Systems 22 (NIPS 2009), 2009.
- [13] M. Bejani, D. Gharavian, and N.M. Charkari, "ANOVA-Based Feature Selection and Multi-Classifer Neural Networks for Audiovisual Emotion Recognition," Neural Computing and Applications, vol. 24, no. 2, pp. 399–412, 2014.
- [14] P.M. Baggenstoss, "Novel Restricted Boltzmann Machines and Deep Belief Networks Applied to Audio Classification," ITG Speech Communication, 2021.
- [15] D. Ayata, Y. Yaslan, and M.E. Kamasak, "Wearable Physiological Sensors for an Emotion-Based Music Recommendation System," IEEE Transactions on Consumer Electronics, 2018.
- [16] X. Yang, Q. Chen, S. Zhou, and X. Wang, "Automatic Music Genre Classification Using Deep Belief Networks," InterSpeech 2011.
- [17] D. Ayata, Y. Yaslan, and M.E. Kamasak, "Emotion-Based Music Recommendation Using Wearable Physiological Sensors," IEEE Transactions on Consumer Electronics, 2018.
- [18] M. Scarpiniti, F. Colasante, S. Di Tanna, M. Ciancia, Y.C. Lee, and A. Uncini, "Audio Classification for Construction-Site Monitoring Using a Deep Belief Network Approach," Expert Systems with Applications (Elsevier), 2021.
- [19] A. Dureha, "A Precise Algorithm for Creating a Playlist from Facial Expressions," IJCA, 2014.
- [20] A. Lehtiniemi and J. Holm, "Employing Animated Mood Images for Music Recommendation," 2012 16th International Conference on Information Visualisation, 2012.