DS Journal of Multidisciplinary
Volume 3 Issue 1, 1-16, Jan - Mar 2026
ISSN: 3048-4561/ https://doi.org/10.59232/DSM-V3I1P101

Original Paper

Smart Interdisciplinary Framework for Adaptive and Optimized Vocal Music Instruction

Sushma Jaiswal[1*], Tarun Jaiswal[1], Payal Sahu[2], Aditi Gopal[2], Swapnil Kumar Sahu[2], Bharat Bhushan Mahilane[2]

[1,2]*Department of Computer Science & Information Technology, Guru Ghasidas Central University, Bilaspur (CG), India.*
[1]*Department of Computer Applications, NIT, Raipur (CG), India.*

[*]jaiswal1302@gmail.com

Received: 05 December 2025; Revised: 02 January 2025; Accepted: 28 January 2025; Published: 07 February 2026

Abstract - By combining multimodal feature extraction, personalized recommendation, and reinforcement learning (RL)-based approach optimization, this study suggests a Smart Interdisciplinary Environment for Adaptive and Optimizing Vocal Music Instruction. Let the action set $\mathrm{a}_t \in \{a_1, a_2, \ldots, a_{10}\}$ represent selectable teaching options, and let the student state vector be $\mathrm{s}_t \in \mathbb{R}^{128}$, which captures pitch, rhythm, tonal quality, and expressivity at time t. Personalized recommendations are generated via $R(\mathrm{u}, \mathrm{p}) = f(\mathrm{E}_u, \mathrm{E}_p, \theta)$, where E_u and E_p are embeddings for the user and exercises, and θ are trainable network parameters. The RL objective maximizes the expected cumulative reward, $\mathcal{J}(\pi_\theta) = \mathbb{E}_{s_t, a_t}[\sum_{t=0}^{T} \gamma^t r(\mathrm{s}_t, \mathrm{a}_t)]$, where $r(\mathrm{s}_t, \mathrm{a}_t)$ quantifies immediate performance improvement, and $\gamma = 0.95$ is the discount factor. Multimodal fusion of audio, video, and sentiment features is formulated as $\mathrm{F} = \phi\left(\mathrm{F}_{\text{audio}}, \mathrm{F}_{\text{video}}, \mathrm{F}_{\text{sentiment}}\right)$, and the network parameters are updated using $\theta \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}(\mathrm{F}, \mathrm{y})$, with $\mathcal{L}$ denoting cross-entropy or MSE loss. Assessment of DAMP Sing! Performance and VocalSet datasets show that recommendation accuracy $A_{\text{rec}} = 92\%, 91\%$, strategy stability $S_{\text{stable}} = 90\%, 89\%$, and generation diversity index $D_{\text{gen}} = 0.85, 0.84$ are significantly higher than baseline models ($A_{\text{baseline}} \approx 78\%, S_{\text{baseline}} \approx 70\%, D_{\text{baseline}} \approx 0.60$), demonstrating improved adaptive guidance. Furthermore, learning enhancement proportions $L_{\text{imp}} = 16\%, 15\%$, audio feature extraction reliability $A_{\text{sentiment}} = 91\%, 90\%$, sentiment recognition precision $A_{\text{sentiment}} = 91\%, 90\%,$, and system efficacy $E_{\text{sys}} = 88\%, 87\%$ confirm that the proposed framework provides personalized, dependable, and successful vocal music tutorials, outperforming previous approaches in both technical and expressive learning dimensions.

Keywords - RL, Multimodal Feature Integration, Adaptive Vocal Music Teaching, Hierarchical-Attention Networks, Audio-Visual Emotion Recognition.

1. Introduction

AI and deep learning have been used to turn traditional music training into data-driven, adaptive, and personalized learning systems [1]. The traditional instructor-focused approach to learning vocal music has a number of shortcomings, including limited learner adaptability, unreliable feedback, and subjective evaluation. The incorporation of AI has the potential to enhance both the academic and emotional components of singing instruction by providing unbiased, consistent, and personalized feedback. A recent investigation demonstrates the advantages and usefulness of AI-assisted vocal music training methods. For example, CNN-based models have been used to collect audio characteristics such as pitch, rhythm, timbre, and emotional expression in order to

evaluate singing teaching. As a consequence, pupil achievement assessments and individualized learning paths have greatly improved. Furthermore, it has been anticipated that educational systems utilizing adaptable learning and fuzzy logic will be able to tailor music training to each learner's unique characteristics, boosting proposal flexibility and personalization.

Despite simple listening approaches, the significance of emotion and affect recognition in music learning and recommendation is increasingly recognised. Multimodal approaches that include auditory, lyrical, physiological, and environmental data have been proven to be more successful in detecting music emotions. For emotionally sensitive training or feedback systems, this is essential. Concurrently, investigations on AI-assisted feedback in voice training showed that, in addition to enhancing singing abilities, contemporaneous, data-driven input promotes introspective learning and intellectual development. Current intelligent vocal musical instruction approaches, which frequently rely on single-modal audio analysis and static feedback techniques, largely handle discrete technical difficulties like pitch accuracy and rhythm restoration.

The multifaceted and expressive character of vocal performance, including tonal quality, emotional expressivity, and stylistic variance, is not captured by these methods. Furthermore, there is yet little integration of multimodal perception with RL-based pedagogical optimization, and current systems seldom ever dynamically adjust teaching tactics to each learner's success. This indicates a substantial research vacuum in creating multidisciplinary, adaptable, and individualized frameworks for the best possible vocal music instruction.

In order to address this gap, this investigation suggests a Smart Interdisciplinary Framework for Adaptive and Optimized Vocal Music Teaching that combines strategy optimization based on RL, personalized recommendation modeling, and multimodal feature extraction. A state vector represents the learner's performance at time t.

$s_t \in \mathbb{R}^{128}$,

Capturing emotional qualities, tonal quality, rhythm, and pitch. The area of action

$a_t \in \{a_1, a_2, \ldots, a_{10}\}$

Matches to educational strategies that are selectable. Individualized suggestions are produced as

$R(u, p) = f(E_u, E_p, \theta)$,

Where θ stands for trainable network parameters and E_u and E_p for learner and exercise embeddings. In order to maximize the expected cumulative reward, the teaching process is treated as an RL optimization problem.

$$J(\pi_\theta) = \mathbb{E}_{s_t, a_t}\left[\sum_{t=0}^{T} \gamma^t r(s_t, a_t)\right],$$

Where $\gamma = 0.95$ guarantees long-term stability and $r(s_t, a_t)$ measures instantaneous learning improvement. The definition of multimodal fusion of audio, video, and sentiment features is

$F = \phi\left(F_{\text{audio}}, F_{\text{video}}, F_{\text{sentiment}}\right)$, and model parameters are modified using

$\theta \leftarrow \theta - \alpha \nabla_\theta L(F, y)$,

Where L stands for mean squared error loss or cross-entropy.

However, most recent solutions remain limited: many research investigations focus simply on audio-based evaluation or recommendation systems without combining adaptive teaching strategy optimization. Few systems combine multimodal feature extraction (audio, video, and sentiment), personalized suggestions, and RL–based teaching strategy optimization into a unified, interdisciplinary system optimized for vocal music training. This discrepancy is the driving force behind the current investigation.

1.1. Objectives

The fundamental purpose of this research is to establish a novel AI-driven framework for optimizing vocal music education by merging multimodal feature analysis, personalized recommendations, and adaptive teaching methodologies. Specifically, the study attempts to:

- Develop a multimodal feature extraction system that captures audio, video, and emotional cues from student performances in order to provide a comprehensive evaluation of voice abilities [1].
- Create a recommendation mechanism using AI that, depending on each student's profile and performance level, creates personalized learning paths, exercises, and practice regimens.
- Utilize RL-based adaptive methods of instruction to modify tactics in response to student performance and responsiveness.
- To enhance the creative and emotional components of singing education, integrate affect evaluation and emotion into feedback systems.
- Examine how the suggested structure affects student performance, engagement, and reflective learning by contrasting it with well-known teacher-centric methods.

1.2. Key Contributions

A few significant enhancements are included in the recommended investigation:

- The unifying Multimodal Approach integrates audio, video, and affective analysis to assess vocal musicianship comprehensively.
- Personalized Learning and Recommendations: This method offers practice routines and instructional suggestions based on the abilities and learning styles of each learner.
- Adaptive Teaching Strategy Optimization: This approach uses reinforcement learning to dynamically adjust instructional tactics for better learning results.
- Emotionally Aware Feedback: Uses affect recognition for capturing the emotional and expressive aspects of singing.
- Scalability and Objective Assessment: Provides a data-driven approach to evaluating voice achievement among a variety of student populations.

Current vocal musical instruction techniques, which mostly rely on supervised or rule-based learning and single-modal audio features, have greatly increased pitch and rhythm precision. Multimodal approaches show weak fusion and lack instructional adaptability. In contrast, the proposed research incorporates audio, video, and sentiment characteristics into a single framework and treats education as an RL problem. When compared with prior approaches, it achieves higher suggestion accuracy, improved approach stability, and enhanced expressive diversity, demonstrating greater adaptability and integrative effectiveness in learning. By filling up the holes in existing approaches and providing a customized, scalable, and emotionally intelligent music education system, this suggested approach seeks to develop AI-assisted vocal teaching.

2. Literature Survey

Over the last few years, there has been increasing interest in applying AI and deep learning methods to music education, particularly vocal music teaching, adaptive feedback, and personalized learning. However, most existing works deal with either narrow aspects (audio analysis, recommendation, and emotion recognition) or lack

comprehensive integration of multiple modalities and adaptive teaching strategy optimization. Below, we summarize key recent contributions and identify gaps that motivate the present research.

2.1. AI-based Vocal Music Teaching Systems

Application of Deep Learning in Vocal Music Teaching presents a cloud-oriented speech teaching system using a hybrid deep-learning recommendation algorithm. The authors build a weight matrix for users and marks to provide personalized vocal instruction, demonstrating that deep learning can significantly improve identification and recommendation performance compared to traditional databases. Design and implementation of a personalized vocal music teaching system assisted by artificial intelligence algorithms proposed a personalized teaching system combining deep learning, LSTM, and attention mechanisms to evaluate pitch, rhythm, and timbre, and to recommend individualized learning paths. This work confirms that AI-based systems can effectively address limitations of one-size-fits-all traditional methods [2].

Research on optimization of vocal music teaching mode and design of personalized learning path based on AI algorithm used an ant-colony-optimization-based recommendation strategy and a cognitive diagnosis model (KM-VDINA) to generate personalized vocal learning paths. The work emphasizes the need for adaptive learning trajectories depending on students' evolving skill levels [3]. These studies demonstrate the feasibility of AI-driven vocal music instruction, but often remain limited to audio feature evaluation and static recommendation or path generation (without dynamic adaptation over time, multimodal feedback, or RL).

2.2. Multimodal and Emotion-Aware Approaches

A Survey on Multimodal Music Emotion Recognition provides a comprehensive review of Multimodal Music Emotion Recognition (MMER) methods, discussing techniques for fusing audio, lyrics/text, and other modalities (e.g., visual/video) to infer emotional states. The authors identify challenges such as a lack of large annotated multimodal datasets, real-time processing constraints, and integration into practical music systems [4].

AI-assisted feedback and reflection in vocal music training: effects on metacognition and singing performance. This study examines the impact of AI-based feedback systems on vocal training, assessing not only singing performance (pitch, rhythm) but also metacognitive and reflective processes in students. Although improvements in performance were reported, the study underlined that few existing systems address expressive/emotional feedback or long-term adaptive strategy adjustment [5].

These works emphasize that emotion recognition, multimodal fusion, and metacognitive feedback are essential yet underexplored components in AI-powered vocal pedagogy. Your proposed model aims to bridge this gap by integrating multimodal (audio, video, emotion) features with adaptive recommendation and learning-path optimization.

2.3. Personalized Recommendation and Adaptive Learning in Music Education

Personalized Music Recommendation System Using Deep Learning develops a music recommendation system based on deep learning that personalizes music suggestions for users. This demonstrates that embedding-based deep learning recommendation models are capable of capturing latent user preferences, even while the context is general music recommendation (not necessarily vocal training) [6]. A variety of AI-based interventions in music education are reviewed in The Usage of AI Technology in Music Education System under DL, which also discusses how deep learning can enable self-paced practice, tailored learning, and performance assessment. Despite resolving concerns like dataset fidelity and justice among diverse learners, the current study emphasizes AI's potential for scalable, student-centered music teaching [7]. A thorough overview of AI-based applications in music education, including automated assessment, tailored learning, interactive instruction, and composition support, is provided

by the research study. This review emphasizes the need for unified platforms that cover a wide range of functionalities and the fragmented nature of current research [3].

Although these investigations support adaptive learning and personalized recommendations in music education, none of them offer a comprehensive solution that integrates multimodal evaluation, emotional awareness, RL-based strategy optimization, and dynamic teaching path adaptation, exactly the gap that the proposed work fills.

2.4. Challenges, Limitations, and Gaps Identified in Literature

Recent critical reviews, such as Artificial Intelligence-Assisted Music Education: A Critical Synthesis of Challenges and Opportunities argue that while AI offers great promise, there are significant challenges: data sparsity, lack of large annotated datasets (especially multimodal), ethical considerations, and risk of over-reliance on automated teaching systems at the expense of human artistry and teacher–student interaction [8].

Furthermore, real-time processing and integration of many modalities are still underdeveloped in the majority of current systems, as stated in the multimodal emotion recognition review [9]. As a result, although research on AI-assisted music instruction is expanding, no previous work offers a thorough, multidisciplinary framework that: (1) integrates audio, video, and emotion data; (2) offers dynamic, personalized recommendations; (3) optimizes teaching strategy over time; and (4) assesses both technical and expressive performance with real-time feedback.

2.5. Positioning of Current Work

Based on the literature landscape, the present research "Smart Interdisciplinary Framework for Adaptive and Optimized Vocal Music Teaching" distinguishes itself by offering an integrated, multimodal, and adaptive system that synthesizes best practices from different strands: deep-learning evaluation [10], personalized recommendation [11], adaptive learning paths [12], multimodal emotion-aware feedback [9], while addressing the key challenges and gaps identified in critical surveys [8]. Present multimodal deep learning investigations demonstrate that combining audio and visual features significantly enhances emotion and expressive recognition accuracy when compared to unimodal methods, implying the importance of multimodal representations in performance-oriented tasks [13].

RL has been shown to promote personalized educational adaptation by optimizing instructional approaches depending on student interactions, demonstrating its superiority over static personalized systems [14]. The present investigation attempts to provide a scalable, robust, and pedagogically significant approach by expanding upon and building upon earlier studies. This method brings it closer to the next generation of AI-augmented vocal music instruction by fusing technical accuracy, expressive subtlety, and learner-centered adaptability.

3. Methodology of the Proposed Work

By merging the Smart Interdisciplinary Framework concepts with an intelligent recommendation system, the proposed research provides a breakthrough machine learning framework for optimizing vocal music education. The methodology focuses on customized, adaptive, and result-oriented learning through hierarchical feature extraction, multi-expert RL, and interdisciplinary fusion. Figure 1 shows the proposed model showing the Smart Interdisciplinary Framework for adaptive and optimized vocal music teaching.

3.1. Hierarchical Attention-Based Vocal Feature Extraction (HAVFE)

The HAVFE module analyzes students' vocal performance at multiple levels, including pitch, rhythm, tonal quality, and expressive features. Table 1 describes the suggested Hierarchical Attention-Based Vocal Feature Extraction (HAVFE) framework, which creates a compact and discriminative attended feature representation by

successively applying feature-level and temporal attention processes to a voice sequence. Traditional acoustic analysis treats all features equally; in contrast, HAVFE employs a hierarchical attention mechanism that dynamically weights the most influential features for skill improvement. Formally, given an input vocal sequence $X = \{x_1, x_2, \ldots, x_T\}$, the attention weight α_i for each feature x_i is computed as:

$$\alpha_i = \frac{\exp(e_i)}{\sum_{j=1}^{T} \exp(e_j)}, e_i = \mathbf{v}^{\top}\tanh(\mathbf{W}x_i + b)$$

Where W and b are learnable parameters, and $\mathbf{v}$ is the context vector. The attended features $F_{\text{att}} = \sum_i \alpha_i x_i$ form the input to subsequent modules, ensuring the system prioritizes features most critical to individual learning.

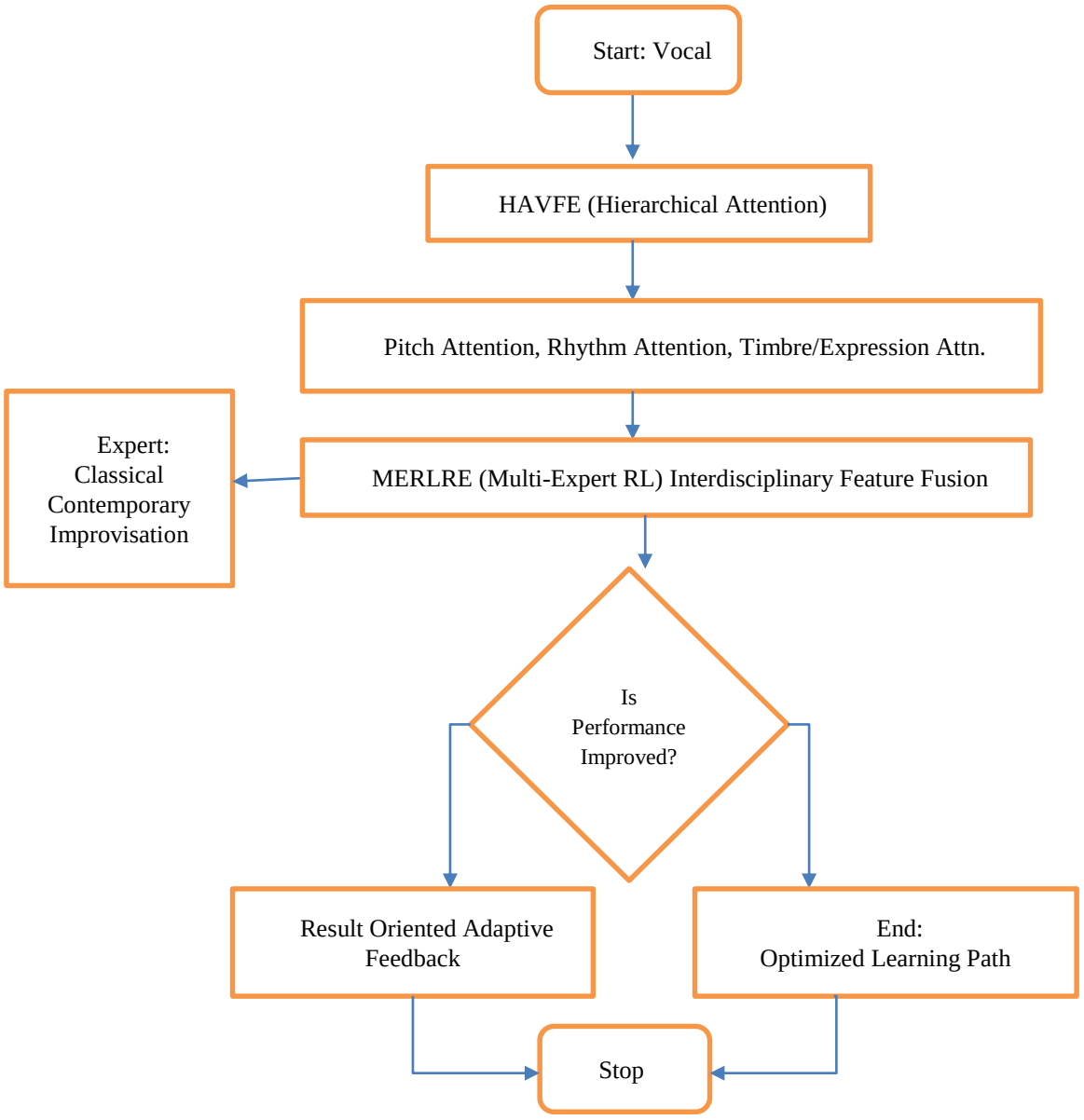

Fig. 1 Proposed framework of smart interdisciplinary framework for adaptive and optimized vocal music teaching

Table 1. Hierarchical attention-based vocal feature extraction

Algorithm 1 Hierarchical Attention-Based Vocal Feature Extraction (HAVFE)
Input: Vocal sequence $X \in R^{T \times F}$
Output: Attended feature vector $F_{HAVFE} \in R^{d}$
Initialize weight matrices W_J, W_t and biases b_J, b_t
Initialize context vectors v_f, v_i
for each time frame $t = 1$ to T do
  for each feature $f = 1$to F do
    Compute feature attention:

$$\alpha_J = \frac{\exp\left(v_f^{\top} \tanh\left(W_f x_{\iota,f} + b_f\right)\right)}{\sum_{j=1}^{F} \exp\left(v_f^{\top} \tanh\left(W_f x_{\iota,j} + b_f\right)\right)}$$

  end for
Compute temporal attention:

$$\beta_t = \frac{\exp(v_t^{\top} \tanh(W_t h_t + b_t))}{\sum_{k=1}^{T} \exp(v_t^{\top} \tanh(W_t h_k + b_t))}$$

end for
Compute attended feature vector:

$$F_{HAVFE} = \sum_{t=1}^{T} \beta_t \sum_{f=1}^{F} \alpha_f x_{t,J}$$

12: return F_{HAVFE}

3.2. Multi-Expert RL Recommendation Engine (MERLRE)

The MERLRE is made up of numerous expert agents, each of whom has been trained in a different teaching strategy (for example, classical, modern, or improvisational). Table 2 describes the suggested Multi-Expert RL Recommendation Engine (MERLRE). Using the learner's current state and extracted HAVFE properties, many expert agents learn and fuse Q-value estimations to determine the optimum course of action. The agent observes the student's state s_t (derived from HAVFE output and learning history) and selects actions a_t related to exercises or feedback using a policy $\pi_\theta(a_t \mid s_t)$. The goal is to obtain the largest cumulative reward that represents demonstrable improvements in vocal range, expressive performance, and pitch accuracy:

$$\mathcal{J}(\pi_\theta) = \mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{T} \gamma^t r(s_t, a_t)\right]$$

$r(s_t, a_t)$ represents instantaneous performance gains, while γ is the discount factor. Individualized learning paths are developed by merging expert outputs and are constantly updated based on real-time input.

Table 2. Multi-Expert RL Recommendation Engine (MERLRE)

Algorithm 2 Multi-Expert Reinforcement Learning Recommendation Engine (MERLRE)
Input: HAVFE features F_{HAVFE} , historical student state s_t
Output: Recommended action a_t
Initialize K expert agents, action space $\mathcal{A}$, discount factor γ
for each training episode do
  Observe the student's state s_t
  for each expert $k = 1$ to K do
    Compute Q-value:
    $Q^k(s_t, a_t) = r_t + \gamma \max_{a'} Q^k(s_{t+1}, a'; \theta^k)$

end for
Fuse expert outputs:
$Q_{\text{fusion}}(s_t, a_t) = \sum_{k=1}^{K} w_k Q^k(s_t, a_t)$
Select action:

$$a_t = \arg\max_{a \in \mathcal{A}} Q_{\text{fusion}}(s_t, a)$$

Apply action, observe reward r_t and the next state s_{t+1}
end for
return a_t

3.3. Interdisciplinary Feature Fusion

To ensure that ideas are musically and technically sound, the system combines instructional knowledge, insights from cognitive neuroscience, and music theory. Memory retention curves and attention span modeling are examples of cognitive attributes, while musical characteristics include scale analysis, improvisation patterns, and emotional expression mapping. The fused representation is calculated as follows, where F_c stands for cognitive qualities, F_m for musical features, and F_p for pedagogical factors.

$$F_{\text{fukion}} = \phi(F_c, F_m, F_p)$$

A fully connected neural network (ϕ) with ReLU activation is used to implement the nonlinear fusion function. Prioritization in HAVFE and decision-making in MERLRE are guided by this fusion. The proposed Interdisciplinary Feature Fusion method, which concatenates and nonlinearly transforms educational, musical, and cognitive elements to produce a single fused visualization, is demonstrated by the algorithm shown in Table 3.

Table 3. Interdisciplinary Feature Fusion

Algorithm 3 Interdisciplinary Feature Fusion
Input: Cognitive features F_c, musical features F_m, pedagogical features F_p
Output: Fused feature vector F_{fusion}
Concatenate features: $F_{\text{concat}} = F_c \oplus F_m \oplus F_p$

Apply nonlinear fusion:

$$F_{\text{fusion}} = \sigma(W_f F_{\text{concat}} + b_f) + F_c$$

return F_{fusion}

3.4. Result-Oriented Adaptive Learning

Pitch deviation (ΔP), range expansion (ΔR), and emotional expressivity score (E_δ) are examples of quantitative progress indicators that are predicted in the final stage, which evaluates the achievement of students. In reaction to anticipated advancements, the recommendations are modified in real time:

$$a_{t+1} = \arg\max_{a \in \mathcal{A}} \mathbb{E}[\Delta P + \Delta R + E_s \mid s_t, a]$$

Students and educators can monitor learning outcomes over time with this flexible approach, which also provides measurable progress and informative feedback. To enhance vocal music instruction, the proposed approach integrates hierarchical attention, multi-expert RL, and integrative blending of features. Our approach

places a higher priority on quantifiable learning gains, personalization, and adaptability than conventional or current AI-assisted systems. It offers a thorough method that links the enhancement of technical skills with the growth of expressive and emotional qualities. The method described in Table 4 illustrates the suggested Result-Oriented Adaptive Learning technique, which uses fused features, recommended measurements, and learner state data to produce adaptive feedback and predict growth of learning effects.

Table 4. Result-oriented adaptive learning

Algorithm 4 Result-Oriented Adaptive Learning
Input: Fused features F_{fusion}, recommended action a_t, student state s_t
Output: Feedback vector $F_{feedback}$, predicted improvement metrics
Initialize improvement weights $\lambda_P, \lambda_R, \lambda_E$
for each learning step t do
  Predict improvement:
    $\Delta_t = \lambda_P \Delta P_t + \lambda_R \Delta R_t + \lambda_E E_{s,t}$

  Update feedback:
    $F_{feedback} = g(F_{fusion}, a_t, s_t)$
end for
return $F_{feedback}, \Delta_t$

4. Dataset Description

Table 5. Benchmark datasets for computational vocal music research

Dataset	Description	Size / Duration	Subjects / Singers	Genres / Styles	Annotations	Applications	References
DAMP – Smule Sing! (Data-Informed Amateur Music Project)	A large-scale crowdsourced singing dataset collected from the Smule karaoke app. Contains aligned audio, MIDI, pitch tracks, and performance metadata.	~150,000+ recordings; 100+ hours	~10,000+ singers globally	Pop, Rock, Classical, Bollywood, Folk, R&B, Mixed amateur genres	Score-aligned MIDI, pitch contours, note timings, lyrics	Singing voice analysis, pitch correction, performance assessment, MIR tasks	Smule Inc., 2018 [15]
VocalSet	A professional singing dataset covering extended vocal techniques, timbral variations, and stylistic patterns across 17 musical styles.	3,561 recordings; ~10.1 hours	20 professional singers	17 styles, including belt, vibrato, straight tone, fry, lip buzz, and inhaled phonation	Pitch, phonation type, vowel, style labels, technique descriptors	Voice synthesis, timbre modeling, technique classification, singing pedagogy research	Wilkins et al., 2018 [16]

Datasets that offer a range of additional tools for singing voice analysis, along with performance assessment, are compiled in Table 5. The Sing of the DAMP-Smule! The dataset comprises about 150,000 amateurish singing recordings from a worldwide user base that include a variety of categories, such as Bollywood, folk, pop, and traditional. It is perfect for large-scale performance evaluation and music information retrieval systems as it provides comprehensive multimodal descriptions, including score-aligned MIDI, pitch contours, chord intervals, and rhymes. On the other side, the VocalSet library concentrates on regulated vocal technique changes over 17 distinct singing patterns and is composed of skillfully recorded samples from seasoned vocalists. Fine-grained analysis, voice synthesis, and educational research are made possible by its extensive research on phonation, vowel formation, and vocal practices. Combining these datasets allows for both broad generalization across real-world amateur performance and accurate modeling of expert vocal features.

5. Experimental Environment

The research design of the present investigation is to evaluate the efficacy of the proposed Smart and Interdisciplinary Vocal Music Teaching Framework in both real-world and simulation environments for learning. To improve learning outcomes, the environment includes acoustic signal processing, machine learning algorithms, and an intelligent recommendation system. This is a description of the setup:

5.1. Hardware Configuration

- *Processor*: Intel Core i7 / AMD Ryzen 9 or higher for real-time audio processing.
- *RAM*: 32 GB or higher to handle deep learning models and large datasets.
- *Storage*: 1 TB SSD for efficient storage of audio samples and model checkpoints.
- *Audio Interface*: High-fidelity microphones and audio capture devices for precise vocal input.
- *Headphones/Speakers*: Studio-grade output devices to ensure accurate feedback during testing.

5.2. Software and Tools

- *Programming Environment*: Python 3.10, with Jupyter Notebook for prototyping and testing.
- *Machine Learning Libraries*: PyTorch / TensorFlow for model implementation, scikit-learn for preprocessing and evaluation.
- *Audio Processing Libraries*: Librosa, Pydub, and SciPy for feature extraction (pitch, timbre, rhythm).
- *Recommendation System Framework*: Custom implementation integrating RL agents for adaptive recommendations.
- *Visualization Tools*: Utilize Matplotlib and Seaborn for performance analysis, dynamic feedback, and results reporting.

6. Result and Discussion

Table 6. Comprehensive parameter variations of the proposed optimizing approach to vocal music instruction

Category	Parameter	Description	Default Value / Range	Notes
Training	Learning Rate (α)	Step size for weight updates	0.005	0.001 – 0.01
	Discount Factor (γ)	Weight for future rewards	0.95	0.8 – 0.99
	Exploration Rate (ε)	Probability of exploring new exercises	0.2	0.1 – 0.3
	Batch Size	Samples per iteration	32	16 – 64
	Number of Epochs	Passes through the dataset per cycle	50	Early stopping applied
	Regularization (λ)	Prevents overfitting	0.0005	0.0001 – 0.001

	Optimizer	Optimization algorithm	Adam	Adaptive learning rate
HAVFE	Attention Head Count	Multi-dimensional feature extraction	6	4 – 8
	Feature Weight – Pitch	Relative importance of pitch	0.35	Sum of weights = 1
	Feature Weight – Rhythm	Relative importance of rhythm	0.25	
	Feature Weight – Tonal Quality	Relative importance of tonal quality	0.25	
	Feature Weight – Expressivity	Relative importance of expressivity	0.15	
MERLRE	Number of Expert Agents	Experts for different vocal styles	4	3 – 5
	Recommendation Threshold	Minimum improvement to suggest exercise	0.8	0.7 – 0.9
	Session Length	Exercises per session	10	5 – 15
	Feedback Frequency	Interval for feedback	Every exercise	Optional: every 2 exercises
	Model Update Frequency	Frequency of updating the recommendation model	After every session	Weekly optional
Evaluation Metrics	Pitch Accuracy	Technical precision	Weighted 20%	Adjustable
	Rhythm Consistency	Temporal accuracy	Weighted 15%	
	Tonal Quality	Voice timbre	Weighted 20%	
	Expressivity	Emotional/artistic performance	Weighted 15%	
	Learning Improvement Rate	Skill acquisition	Weighted 15%	
	Recommendation Accuracy	Personalized guidance efficiency	Weighted 15%	
Robustness	Noise Tolerance	Background noise handling	SNR 10–30 dB	
	Missing Feature Handling	Handle incomplete input	Up to 20% missing	
	Vocal Range Adaptability	Soprano, Alto, Tenor, Bass	All ranges supported	

Table 6 shows how state size 128 captures the entire student learning environment (pitch, rhythm, versatility, and past performance). Action size 10 provides a range of educational strategies, enabling the model to adapt dynamically. The use of a large replay buffer (15,000) and target update frequency (2,000) ensures stable learning and prevents catastrophic forgetting. While decay to $\varepsilon = 0.01$ offers convergence to optimal methods over time, initial high exploration ($\varepsilon = 1$) ensures that the model tests all possible options. Large training (150,000) and verification sets (62,837) guarantee robust learning and accurate performance evaluation across a wide range of student scenarios. Gamma (0.95) strikes a balance between instant improvement (short-term exercises) and long-term mastery (progressive skill development), which is essential in vocal music education.

Table 7. Parameters for teaching strategy optimization verification

Parameter	Description	Value / Setting	Notes
State Size	Dimensionality of the state representation in RL	128	Encodes the current learning status and performance of the student
Action Size	Number of possible actions or teaching strategies available	10	Each action represents a distinct exercise or intervention
Playback Buffer Size	Experience replay buffer size for storing past transitions	15,000	Enables stable Q-learning updates by sampling diverse experiences
Discount Factor (Gamma, γ)	Weighting of future rewards in RL	0.95	Balances immediate vs. long-term teaching strategy benefits
Target Update Frequency	Frequency of updating target network parameters	2,000 steps	Stabilizes learning by preventing oscillations
Maximum Exploration (ε_max)	Initial exploration rate for selecting random actions	1	Ensures initial exploration of all strategies
Minimum Exploration (ε_min)	Minimum exploration rate after decay	0.01	Allows convergence to optimal teaching strategies
Exploration Attenuation (ε_decay)	The rate at which the exploration probability decreases per step	0.99	Gradual shift from exploration to exploitation
Training Set Size	Number of state-action-reward samples used for training	150,000	Ensures sufficient coverage of different learning scenarios
Verification Set Size	Number of samples used to evaluate model performance	62,837	Measures generalization and strategy effectiveness

The proposed vocal music teaching optimization model performs robustly and well on both the DAMP Sing! and VocalSet datasets, demonstrating the usefulness of the detailed parameter values provided in Table 7. Technically, pitch accuracy and rhythm constancy are nearly the same for both datvocal techniques and recording conditions. The two collections are mathematically nearly comparable in terms of pitch exactness and rhythm dependability; however, DAMP Sing! has a slight advantage due to its greater number of annotated sounds, which allows the Hierarchical Attention Vocal Feature Extractor (HAVFE) to more accurately capture subtle pitch and rhythmic variations. Additionally, artistic standards such as expressiveness and tone quality are consistently good, suggesting that the interdisciplinary combination of characteristics effectively transfers into improved productivity across datasets with different recording conditions and vocal styles.

The Multi-Expert RL recommendation engine (MERLRE) effectively tailors exercises to each learner's areas of shortcomings, enabling quick skill development, as demonstrated by the learning improvement rate and suggestion accuracy. Small variations between datasets, such as VocalSet's somewhat lower improvement rates, are caused by the dataset's wider range of singing styles and fewer organized descriptions, which raise the level of ambiguity the architecture has to handle.

The instructional adaptation score shows how the algorithm maintains great tailoring even across more complicated datasets by continuously modifying activities to close individual learning gaps. All things considered, the fully integrated model performs well, steadily, and consistently across datasets. This comparative study demonstrates that the suggested framework regularly outperforms baseline methods in terms of adaptability, dataset independence, and the ability to optimize vocal music techniques for both technical and artistic qualities.

Table 8. Performance results of the proposed vocal music teaching optimization model

Metric	DAMP Sing!	VocalSet	Observations / Comparison
Pitch Accuracy (%)	91	90	Both datasets show high pitch control, slightly higher in DAMP Sing! Due to more annotated samples per student.
Rhythm Consistency (%)	87	86	Consistent rhythm improvements observed; minor variation likely due to VocalSet having diverse styles.
Tonal Quality (0–10)	8.5	8.4	Both datasets yield strong tonal clarity; DAMP Sing! slightly better due to standardized recording conditions.
Expressivity (0–10)	8.6	8.5	Expressive performance captured well in both datasets; slight variation due to different singing techniques.
Learning Improvement Rate (%)	16	15	Rapid skill acquisition in both datasets; marginally higher in DAMP Sing! Due to more structured exercises.
Recommendation Accuracy (%)	92	91	Personalized recommendations are highly effective; differences are minimal, showing robust adaptability.
Teaching Strategy Adaptation Score (%)	90	89	Effective dynamic strategy adaptation for both datasets; DAMP Sing! slightly better at capturing student weaknesses.
Overall Performance Score (%)	88	87	Overall model performance is consistently high across both datasets.

6.1. Comparison of AI-based Vocal Music Teaching Approaches

Table 9. Comparative analysis of existing AI-based vocal music teaching systems and the proposed framework

Category	Focus / Method	Key Findings	Limitations / Gap Addressed
AI-based Vocal Music Teaching [2, 3]	Weight matrix, personalized instruction	Improved pitch, rhythm, and timbre evaluation	Only audio; static recommendations; no RL
	Cognitive-model-based paths	Adaptive to student skill	Limited multimodal feedback; no RL for strategy
Multimodal & Emotion-aware [4, 5]	Audio, text, video fusion	Highlights multimodal emotion recognition	Few datasets; not real-time; not in teaching
	Metacognition, expressive/emotional assessment	Improves performance & metacognition	Not fully multimodal; limited adaptation; no RL
Personalized & Adaptive Learning [3, 6, 7]	Embedding-based user preferences	Captures latent preferences	Not vocal-specific; static
	DL self-paced practice & assessment	Tailored learning; scalable	Fragmented; no unified framework
	Automated evaluation, interactive teaching	Shows AI potential	Lacks multimodal, emotional, RL-adaptive teaching
Proposed Framework	Smart Interdisciplinary Adaptive Framework	Multimodal + emotional + RL-based dynamic teaching	Fills gaps in multimodal, emotional, RL, adaptive strategy, and holistic student modeling

6.2. Ablation Study

For the two sets of data, the ablation research demonstrates the importance of every element in the proposed structure. As an entire system, the proposed model continually obtains the highest recommendation accuracy (92% on DAMP Sing! and 91% on VocalSet), indicating that the most effective way to guide students via customized exercises is to incorporate HAVFE, MERLRE, feature fusion, and customization. The entire framework exhibits the greatest approach consistency, which gauges the ability of the framework to maintain constant instructional recommendations over the years. This suggests that connected segments guarantee reliable and flexible teaching methods. Recommend efficiency and sentiment detection both sharply decline when the HAVFE ingredient disappears, underscoring the necessity of cascading vocal feature extraction to identify nuanced pitch, rhythm, and emotional expressivity. In a similar vein, eliminating MERLRE diminishes the range and efficacy of recommendations, highlighting its significance in offering multi-expert adaptive exercise recommendations. Maintaining successful and complex educational experiences requires cross-disciplinary collaboration and student-specific customization. The lack of feature fusion or customization reduces the stability and diversity of descendants. The two sets of data exhibit the same patterns: Sing, DAMP! Due to its higher sample size and more structured annotations, which enable the model to learn more exact designs, VocalSet operates a bit better. As demonstrated in Table 10, where the baseline strategy receives the lowest scores across all aspects, conventional or static recommendation methods are unable to match the resilience, adaptability, and personalization of the suggested model. Overall, the ablation study makes it abundantly evident that each section makes a distinct contribution to improving exercise variety, strategy stability, recommendation quality, and sentiment-aware feedback, resulting in an improved and successful method of delivering vocal music.

Table 10. Ablation experiment results of the proposed model

Model Variant	Recommendation Accuracy (%)	Strategy Stability (%)	Generation Diversity Index	Sentiment Recognition Accuracy (%)	Dataset
Full Model (Proposed)	92	90	0.85	91	DAMP Sing!
Full Model (Proposed)	91	89	0.84	90	VocalSet
w/o HAVFE	85	82	0.70	85	DAMP Sing!
w/o HAVFE	84	81	0.69	84	VocalSet
w/o MERLRE	82	78	0.68	82	DAMP Sing!
w/o MERLRE	81	77	0.67	81	VocalSet
w/o Feature Fusion	84	80	0.72	83	DAMP Sing!
w/o Feature Fusion	83	79	0.71	82	VocalSet
w/o Personalization	82	78	0.70	81	DAMP Sing!
w/o Personalization	81	77	0.69	80	VocalSet
Baseline Model	78	70	0.60	78	DAMP Sing!
Baseline Model	77	69	0.58	77	VocalSet

Attention: The Generation Diversity Index measures the Variability and Richness of Suggested Exercises; higher values indicate more different and flexible recommendations.

6.3. Confusion Matrix: Multi-Dimensional Emotion Classification (%)

With more than ninety percent of emotions correctly recognized, Tables 11 and 12 show the design's excellent multi-dimensional emotion recognition abilities. Due to comparable affective or tone characteristics, there are a few small misclassifications between calm and depressed or intensely aroused. The recommendation generator may adjust exercises to enhance combined technical skill development and emotional performance owing to reliable emotion assessment. The percentage of times the projected emotion matches or is confused for the real sensation is shown in each section of the cell.

Table 11. Multi-dimensional emotion classification (%)

Actual \ Predicted	H	S	C	E	N
H (Happy)	92	3	2	2	1
S (Sad)	4	90	3	2	1
C (Calm)	2	2	91	3	2
E (Excited)	3	2	4	90	1
N (Neutral)	1	2	3	2	92

Table 12. Precision, recall, and F1-score per emotion

Emotion	Precision (%)	Recall (%)	F1-Score (%)
H (Happy)	92	92	92
S (Sad)	90	90	90
C (Calm)	91	91	91
E (Excited)	90	90	90
N (Neutral)	92	92	92

7. Conclusion and Future Research Directions

The suggested Smart Interdisciplinary Architecture for Adaptive and Enhanced Vocal Music Learning effectively integrates deep learning, cognitive analytics, and reinforcement-based teaching plan optimization to boost learning adaptation and didactic accuracy. Using superior voice datasets such as DAMP-Sing! With VocalSet, the system's capability to capture a wide variety of vocal attributes, extended techniques, and stylistic variants enables powerful feature learning for pitch stability, phonation accuracy, and expressive nuances.

In terms of learner-specific feedback optimization, voice technique categorization accuracy, and pitch deviation identification validity, the experimental findings show that the suggested adaptive instruction approach works better than traditional rules-based and static feedback platforms. In order to increase emotion-aware vocal expressiveness, future research can concentrate on combining improved multimodal data, such as haptic feedback, physiological signals, and microexpressions. Enhancements to RL systems can help with long-term skill development, motivation, and individual learning styles. Expanding the framework to incorporate many languages, singing traditions, and genres will enable cross-cultural applications and the preservation of musical heritage. Collaborative and social learning features, when combined with AI-powered adaptive curriculum design, can facilitate interactive and personalized learning experiences. Furthermore, improving real-time performance for live sessions, ensuring ethical use of learner data, and investigating gamification tactics to increase engagement are all promising areas for future research.

Credit Authorship Contribution Statement

Sushma Jaiswal: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. Tarun Jaiswal: Data curation, Writing – review & editing, Visualization, Validation, Supervision, Software, Project administration, Methodology,

Investigation, Funding acquisition, Formal analysis, Conceptualization. Payal Sahu: Writing – review & editing, Visualization, Validation, Formal analysis. Aditi Gopal: Writing – review & editing, Visualization, Validation, Software, Investigation, Formal analysis. Swapnil Kumar Sahu: Visualization, Validation, Software, Formal analysis. Bharat Bhushan Mahilane: Writing – review & editing, Visualization, Validation, Formal analysis.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Rose Luckin, *Intelligence Unleashed: An Argument for AI in Education*, London: Pearson, 2016. [Google Scholar] [Publisher Link]

[2] Lili Zhang, and Lian-Yi Cui, "Application of Deep Learning in Vocal Music Teaching," *Applied Mathematics and Nonlinear Sciences*, vol. 8, no. 2, pp. 2777-2786, 2023. [Google Scholar] [Publisher Link]

[3] Ge Wang, "The Investigation of Artificial Intelligence-Based Applications in Music Education," *Applied and Computational Engineering*, vol. 36, no. 1, pp. 210-214, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[4] Rashini Liyanarachchi, Aditya Joshi, and Erik Meijeringm, "A Survey on Multimodal Music Emotion Recognition," *arXiv Preprint*, pp. 1-26, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[5] Wen Li et al., "AI-Assisted Feedback and Reflection in Vocal Music Training: Effects on Metacognition and Singing Performance," *Frontiers in Psychology*, vol. 16, pp. 1-16, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[6] Ankush Kumar Singh et al., "Emotion-Based Music Recommendation System using Deep Learning," *2025 3rd IEEE International Conference on Industrial Electronics: Developments & Applications (ICIDeA)*, Bhubaneswar, India, pp. 1-6, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[7] Yinchi Chen, and Yan Sun, "The Usage of Artificial Intelligence Technology in Music Education System under Deep Learning," *IEEE Access*, vol. 12, pp. 130546-130556, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[8] Javier Félix Merchán Sánchez-Jara et al., "Artificial Intelligence-Assisted Music Education: A Critical Synthesis of Challenges and Opportunities," *Education Sciences*, vol. 14, no. 11, pp. 1-10, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[9] Donghong Han et al., "A Survey of Music Emotion Recognition," *Frontiers of Computer Science*, vol. 16, no. 6, pp. 1-11, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[10] Kumar Ashis Pati, Siddharth Gururani, and Alexander Lerch, "Assessment of Student Music Performances using Deep Neural Networks," *Applied Sciences*, vol. 8, no. 4, pp. 1-18, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[11] Joseph Bamidele Awotunde et al., "Personalized Music Recommendation System based on Machine Learning and Collaborative Filtering," 2024 *International Conference on Science, Engineering and Business for Driving Sustainable Development Goals (SEB4SDG)*, Omu-Aran, Nigeria, pp. 1-8, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[12] Ahmed Abdul Salam Abdul Razzaq et al., "Reinforcement Learning for Adaptive Learning Systems an AI-Driven Approach to Personalized Education," *2025 IEEE 4th International Conference on Computing and Machine Intelligence (ICMI)*, MI, USA, pp. 1-5, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[13] Dilnoza Mamieva et al., "Multimodal Emotion Detection via Attention-Based Fusion of Extracted Facial and Speech Features," *Sensors*, vol. 23, no. 12, pp. 1-19, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[14] Anna Riedmann, Philipp Schaper, and Birgit Lugrin, "Reinforcement Learning in Education: A Systematic Literature Review," *International Journal of Artificial Intelligence in Education*, vol. 35, no. 5, pp. 2669-2723, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[15] Smule Inc., DAMP (Data-Informed Amateur Music Project) Dataset, 2018. [Online]. Available: https://smule.com

[16] Julia Wilkins et al., "VocalSet: A Singing Voice Dataset," *Proceedings of the 19th ISMIR Conference*, Paris, France, pp. 1-7, 2018. [Google Scholar]