DOI: 10.6977/IJoSI.202512_9(6).0004
Q. Liu & T.I. Tajuddin/Int. J. Systematic Innovation, 9(6), 44-62 (2025)

Artificial intelligence-collaborative folk music composition system based on gesture recognition: A real-time interactive framework integrating computer vision and folk music generation

Qinghao Liu[1], Tazul Izan Tajuddin[1,2]*

[1]Faculty of Music, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia

[2]Institut Seni Kreatif Nusantara, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia

*Corresponding author E-mail: tazulizan@uitm.edu.my

(Received 11 August 2025; Final version received 07 November 2025; Accepted 02 December 2025)

Abstract

Artificial intelligence (AI) and gesture recognition offer new creative possibilities, yet culturally sensitive, real-time systems for gestural folk music composition remain largely undeveloped. This study develops an AI-collaborative folk music composition system that integrates computer vision-based gesture recognition with specialized folk music generation algorithms to create a real-time interactive framework that supports traditional music composition while preserving cultural musical characteristics across multiple folk traditions. The system employs a four-layer architecture encompassing gesture acquisition, computer vision processing, interpretation, and generation layers. A comprehensive dataset of 1,643 folk music compositions from established repositories representing English, American, Irish, and Chinese traditional music (Nottingham Dataset, Irish Traditional Corpus, and self-recorded materials) was curated, supplemented by 6,127 successfully tracked gesture samples collected from 47 participants across 12 folk music gesture categories. The evaluation framework assessed gesture recognition accuracy, cultural authenticity preservation, real-time performance, and collaborative effectiveness through extensive experimental validation. The system achieved robust gesture recognition performance with 88.9% accuracy and 23.4 ms processing latency, while maintaining end-to-end response times of 86.8–91.6 ms during collaborative sessions. Cultural authenticity scores ranged from 7.6 to 8.3 across different regional folk styles, with a user satisfaction rating of 7.8 and a 28% improvement in musical coherence compared to baseline approaches. The framework successfully supports up to eight concurrent users while maintaining sub-100 ms real-time performance requirements. The integrated system successfully demonstrates effective coordination between gesture recognition and folk music generation subsystems, validating the architectural design and optimization strategies for culturally sensitive AI applications across diverse folk music traditions. The validated framework provides a foundation for educational, performance, and cultural preservation applications, contributing methodological insights for multimodal human–AI interaction systems and culturally aware creative technologies applicable to traditional music contexts.

Keywords: Artificial Intelligence-Collaborative Music Composition, Computer Vision, Folk Music Generation, Gesture Recognition, Real-Time Interactive Framework, Traditional Music

1. Introduction

Rapid development of artificial intelligence (AI) has greatly altered many areas of human creation, and one of the promising areas for AI application is musical composition (Hernandez-Olivan & Beltran, 2022). Combining computer vision, gesture recognition, and AI-generated music reveals new possibilities for designing intelligent systems that can interpret and respond based on human creative intention under real-world conditions. This integration would be quite appealing for folk music creation, as it offers opportunities to develop sophisticated human–computer interaction patterns for diverse geographic and ethnic origins of traditional culture, and still maintain its authenticity and complexity. This research

exclusively focuses on designing a framework for composing multicultural folk music that would cover the traditional music of different regions, such as English, American, Irish, and Chinese folk music, taking into consideration that each folk music tradition has some commonalities while also retaining distinctive characteristics.

Recent advancements in AI research have shown remarkable success in many methods, ranging from deep architectures to symbolic music modeling systems (Ji et al., 2023). In this regard, automatic methods of composing music through AI systems remain a broad area of activity, encompassing many computational approaches (Civit et al., 2022). Recently, comprehensive research on AI music composition applications demonstrates that there is a growing number of choices available for incorporating this technology creatively (Chen et al., 2024). Nevertheless, applications of these technologies, despite their great potential for culture preservation and creation, have been rarely explored, especially concerning music composition for folk culture.

These distinctive elements, including its melody, rhythmic components, and expression of emotion, have been shown to be supportive and motivating, not only for the AI system that is attempting to mimic intelligent output, but also for producing more valid output (Sturm & Ben-Tal, 2021). In addition, going beyond its technical applications, its philosophical applications for AI systems within music involve more basic concepts of creating, authoring, and originality of culture itself (Berkowitz, 2024). Text-to-music creation techniques demonstrate ongoing research into the advanced functionality of AI techniques for producing large stylistic output from text itself (Zhao et al., 2025). Furthermore, the future applications of machine learning techniques into music continue to expand into this new, imaginable realm of human–computer interaction and creation (Liang, 2023).

Machine learning and computer vision have kept pace with developments in modern gesture recognition systems, and more powerful tools have emerged for high-quality and interactive music systems (Dalmazzo et al., 2021). There have been outstanding breakthroughs in computer vision methods of musical transcription, and the processing of visual information into symbolic music expression has been achieved effectively (Li et al., 2020). The successful deployment of such gesture recognition capabilities within musical contexts requires systematic integration across multiple computational layers, addressing distinct processing requirements from visual input to musical output. Systems enabling multimodal interaction between human and computer are becoming increasingly able to understand and interpret complex user inputs performed through more than one perceptual channel (Jia et al., 2020). The improvement of performer–audience interaction through technological mediation has become an emerging research area with significant implications for the live musical performance experience (Otsu et al., 2021).

Affective music composition systems face inherent challenges in composing music that is emotionally meaningful to humans in terms of human emotions and cultural background (Dash & Agres, 2024). The application of machine learning in music generation and composition has been successful across diverse music genres and styles (Dawande et al., 2023). Deep network architectures tailored for music generation are becoming increasingly complex to produce higher-quality output (Pricop & Iftene, 2024). Extensive investigation of multimodal interaction interfaces effectively demonstrates the challenges in developing robust systems for human–computer communication (Dritsas et al., 2025), requiring careful coordination among gesture capture, interpretation, and generation subsystems to maintain coherent real-time operation.

Generative AI in music has educational potential and implications for traditional pedagogy, considering how this technology might be integrated (Cheng, 2025). The synthesis of music sound using machine learning techniques has made considerable progress in producing perceptually relevant control of a synthesizer through user input (Roche, 2020). In commercialized musical contexts, the implementation of deep neural networks in music industry processes provides pragmatic evidence of AI technology adoption (Fan, 2022). Moreover, psychological studies of neural processing suggest that musicians are trained to have heightened sensitivity to music-relevant aspects of the musical context, acquired through experience (Hansen et al., 2022).

The appearance of music perception skills in neural networks without human supervision is in line with our hypothesis that the networks might develop a kind of internal understanding of music through learning from various musical pieces (Kim et al., 2024). Deep learning approaches based on music genre identification have made remarkable progress, demonstrating their potential for better style recognition and generation (Yimer et al., 2023). Studies on soundscape features in human–computer interaction contexts have demonstrated the relevance of environmental and contextual aspects when designing musical interfaces that effectively coordinate music-sound with non-music sound (Johansen et al., 2022). Dynamically coping with uncertainty is a continuing problem for machine learning systems implemented for applications that need guaranteed and predictable performance (Kapoor, 2025).

Fine-grained interactive guidance for symbolic music generation is also a major step toward

obtaining full control over the operations of AI music authoring (Zhu et al., 2024). However, despite such achievements, many limitations of current AI music generation systems exist, limiting their application effectiveness in folk music. Many models cannot be sensitive enough to the subtle attributes of folk music traditions and inevitably generate results that are not realistically stylized due to constraints in meaningful cultural expressions. The challenge is exacerbated when considering the real-time nature of interaction required for gesture-based systems, where latency and responsiveness are critical to preserving an effective creative flow.

To effectively address these issues of real-time system performances, especially under uncertain operating settings, sophisticated control techniques from the theoretical framework of non-linear systems, as described by gesture-based human–computer interaction, need to be applied. Recent efforts based on fuzzy control, aiming for practical fixed-time synchronization of fractional-order chaotic systems (Boulkroune et al., 2025), and output-feedback controllers based on projective lag synchronization of uncertain chaotic systems subjected to input non-linearities (Boulkroune et al., 2017), have laid theoretical foundations for addressing interoperability and system-wide time compatibility between complex computational components with diverse time characteristics. These approaches can be collectively considered as offering methodologies to provide practical solutions for time compatibility and gesture synchronization, particularly under variable lighting or sensor noises in gesture recognition algorithms used as input components of gesture-to-music translators.

To address complex interactions between gesture, musical parameters, and generators—aiming for comprehensive theoretical foundations—control structures from neural networks, especially as proposed for uncertain complex dynamical multivariable systems and based on advanced neural network online prediction methods (Zouari et al., 2012), and hierarchical adaptive backstepping methods concerning uncertain single-input single-output (SISO) non-linear systems (Zouari et al., 2013a), which align conceptually with multi-layer processing architectures, are required for gesture-to-music translation.

Gas-compressor systems based on induction motors, utilizing practical implementation and validation through non-linear optimal control methods (Rigatos et al., 2023), and flexible robotic systems based on DC motors—adopting practical implementation through adaptive backstepping control methods (Zouari et al., 2013b)—have demonstrated that complex adaptability techniques can be effectively applied within strict time-constrained responsive systems. These control methods could be integrated into gesture–music systems involving gesture-based folk music generation, potentially offering improved resistance to environmental changes, as well as more accurate gestural expression interpretation, while maintaining cultural integrity required for folk music tradition and transmission.

Gesture recognition and AI-generated music face complex technical challenges, as they involve coordinating multiple tasks across different computer systems. Research on gestural control of AI-generated folk music is currently facing challenges, particularly the need for research that focuses on AI-generated folk music while taking into consideration characteristic features and traditions within different cultures worldwide. By addressing this research gap, the present study develops a comprehensive framework that incorporates advanced computer vision technology and specialized AI-generated folk music algorithms based on multicultural materials of traditional folk music, establishing a system of gestural human–AI cooperation on music creation grounded in Western, Celtic, and Eastern folk music traditions.

2. Methods

2.1. System Architecture Design

To develop the proposed AI-assisted folk music composition system, a complex architectural design is required that effectively incorporates gesture processing functionality and folk music processing functionality into one efficient system that is still fully responsive and culturally compatible with diverse folk music traditions. The system design was formulated to address one of its core challenges, namely, managing multiple computer processing systems operating at different time scales while retaining the natural, free-flowing characteristic typical of folk music performances. To enable a clear understanding of this complex system, Fig. 1 outlines the architectural system design for the proposed system.

Fig. 1 illustrates the system architectural hierarchy, where the gesture acquisition layer interacts solely with computer vision processing modules. Gestural input is bridged via musical parameters by the interpretation layer, and folk music composition based on culturally appropriate folk music is achieved through specialized neural modules in the generation layer. There are four primary processing layers that operate in parallel to optimize latency, system responsiveness, and musical composition quality. In addition, a dependency relationship exists between each of these processing layer components, with input components receiving raw streaming data (e.g., depth sensor images for gesture acquisition, feature vectors for gestural input, musical parameters for musical composition), processing components implementing

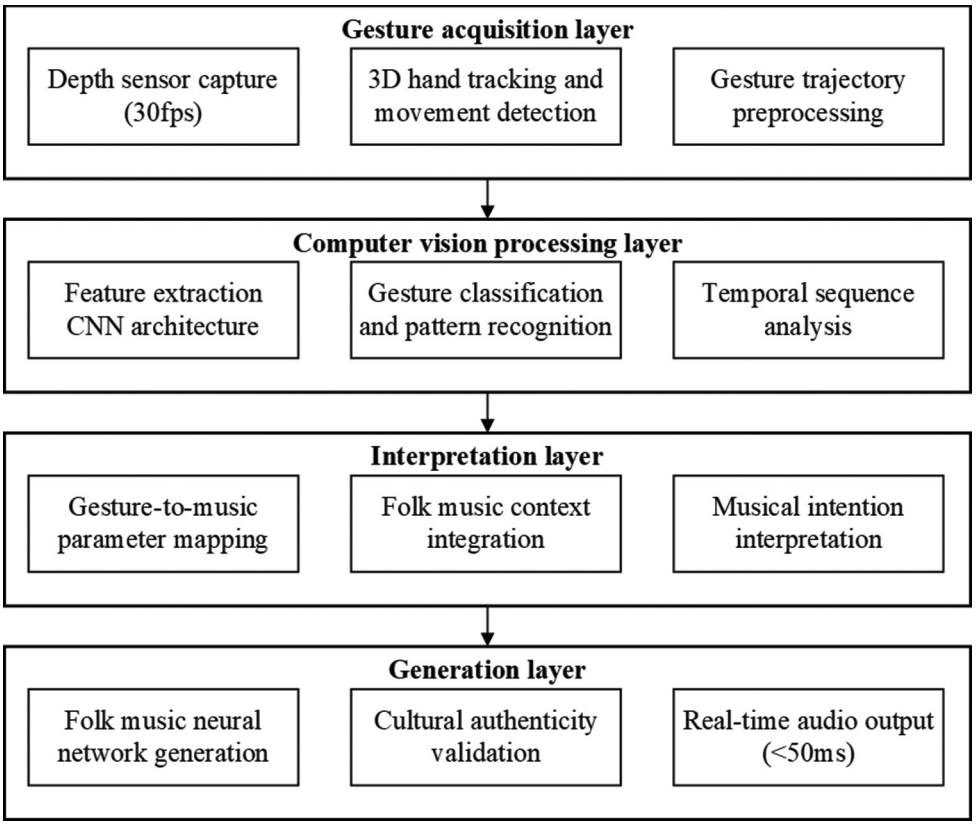

Fig. 1. Multi-layer system architecture for artificial intelligence-collaborative folk music composition
Abbreviations: 3D: Three-dimensional; CNN: Convolutional neural network

operations specific to each respective layer (e.g., spatial–temporal feature extraction for computer vision processing, gestural input processing, and constraint-based musical composition for generation), and output components that verify and pass valid processing output downstream through standardized interfaces for continuous system operation and final musical composition output.

The proposed common architecture compares favorably with other methods due to its ability to incorporate domain-specific parameters of folk music structures and performance practices into its framework, thereby facilitating more informed gesture interpretive processing and more appropriate musical responses based on these considerations (Rezwana & Maher, 2023). To address issues of system operation uncertainty and adaptability, the system design integrates control concepts based on principles of non-linear systems theory, incorporating an adaptive feedback system that continuously monitors and responds based on gesture recognition degrees of confidence and processing delays. This is achieved through decision rules formulated on fuzzy logic and techniques of hierarchical backstepping control, that is, system operations that aggregate performance levels from lower processing scales to higher gesture or co-processing scales over ongoing system operation, with parameters governed by backstepping control methodologies.

A key benefit for developers is the ability to optimize each goal separately, namely, gesture recognition accuracy and music output quality, due to straightforward inter-module communication based on a set of standardized data exchange protocols. All elements of this system operate within the time constraints of their respective stages, ensuring that all operations occur in real time. This applies to gesture processing, which supports up to 30 fps camera capture, and music generation, which maintains a maximum latency of 50 ms. In addition, the systems include facilities for dynamic resource allocation, allowing processing priorities to adjust based on system load and interaction intensity, thereby maintaining stable quality of service across varied system configurations.

2.2. Computer Vision-Based Gesture Recognition and Real-Time Interaction

The computer vision module develops new state-of-the-art deep learning architectures designed for musical gesture recognition and real-time classification in folk music performance. The gesture recognition methodology uses hierarchical feature extraction methods, which exploit spatial–temporal hand

movement patterns to inform secondary models based on convolutional neural networks trained to discriminate between intended musical gestures and unintended body motion observed during creative sessions. To illustrate the complex processing chain that converts raw gestural input into musical parameters, we present a detailed description of the gesture recognition pipeline in Table 1.

Table 1 presents a comprehensive system configuration of gesture recognition, including input data formats, feature extraction and classification algorithms, and processing of time scales necessary for comprehending gestural expressions. This design takes into consideration parameters such as Gaussian spatial smoothing—σ of 1.2, determined by experimental criteria aimed at balancing noise elimination with hand movement detail preservation. It also incorporates a time window width of 15 frames, based on the criterion of processing time below 30 ms, and a prediction threshold of 0.85, chosen to minimize false positives during prolonged interaction. The gestural classification framework captures movement on varied time scales, enabling comprehension of immediate reactive gesture as well as artistic gestural expressions corresponding to musical interpretations that remain cognizant of cultural context.

To facilitate better visualization of the sequence of data flows and interdependencies within the four processing stages outlined in Table 1, Fig. 2 shows a comprehensive flow diagram illustrating the gesture recognition pipeline architecture, including the corresponding data transformations and/or operations at each stage of processing.

Fig. 2 illustrates the step-by-step processing pipeline that generates valid gesture classifications from raw depth sensor input, based on four synchronous processing tasks executed under stringent time restrictions. The total processing latency of 23.4 ms, well below the target value of 30 ms corresponding to the frame rate of 30 fps, demonstrates an optimal pipeline that renders gestural interaction effortless and free of delays. Every processing task is allocated its own processing resources, as indicated in Table 2, and buffer handling ensures loss-free processing during concurrent user interactions, which is critical for gestural sequence processing tasks involving longer durations of collaborative activity.

The 12 folk music gesture types include universal conducting gestures that can be applied across different traditions of tempo, dynamic marking, and boundary marking. Culture-specific expressive gestures typical of a particular folk music tradition—including articulation gestures, expression gestures, and melodic indication gestures—refer to pitch gestalt characteristic of scales typical of that tradition. While some of these gestures can be interpreted based on general musical principles, others exhibit culture-specific nuances unique to certain folk music performances.

Moreover, the respective modules for recognition and feature extraction are enabled using advanced machine learning techniques, allowing the system to adapt through continuous learning, especially based on observation and experience. This enhances its reliability under uncertain environments, particularly when recognizing gestures under continuously

Table 1. Gesture recognition processing pipeline specifications

Component	Parameter	Specification	Unit
Input data format	Depth resolution	640×480	Pixels
	Frame rate	30	Fps
	Data type	16-bit depth	-
	Detection range	0.5–4.5	Meters
Feature extraction	Hand keypoints	21	Points
	Feature vector dimension	126	Dimensions
	Temporal window size	15	Frames
	Spatial smoothing filter	Gaussian $\sigma = 1.2$	-
Classification algorithm	CNN layers	3 convolutional+2 FC	Layers
	LSTM hidden units	128	Units
	Gesture classes	12 folk music gestures	Classes
	Model parameters	2.4 M	Parameters
Temporal processing	Processing latency	23	ms
	Buffer size	8	Frames
	Prediction confidence threshold	0.85	-
	Gesture sequence length	0.5–3.0	s

Abbreviations: CNN: Convolutional neural network; FC: Fully-connected; LSTM: Long short-term memory.

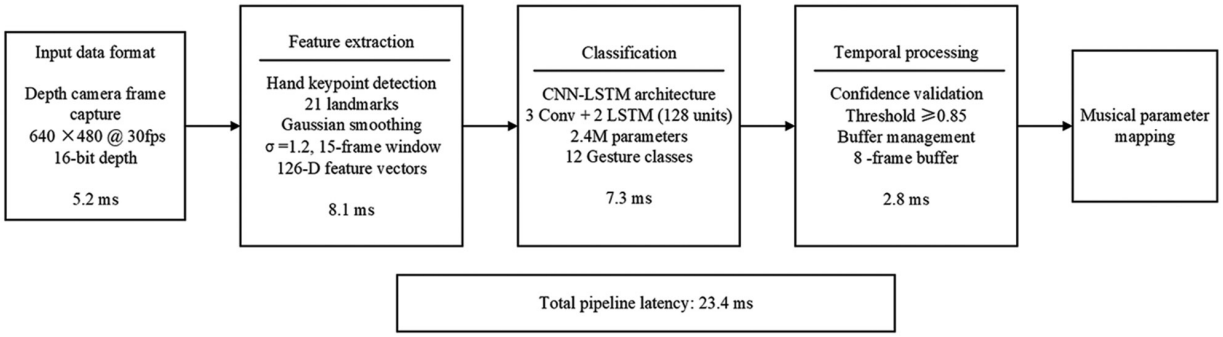

Fig. 2. Gesture recognition pipeline flow diagram
Abbreviations: CNN: Convolutional neural network; Conv: Convolutional; LSTM: Long short-term memory

Table 2. System integration and real-time optimization configuration

Integration component	Parameter	Specification	Notes
Inter-module communication	Data transfer protocol	UDP with error correction	Low-latency priority
	Message queue size	256 buffers	Ring buffer
	Latency requirement	<10 ms	Design target
	Buffer overflow handling	Drop the oldest policy	-
Temporal synchronization	Master clock frequency	48 kHz	Audio sample rate
	Sync tolerance	Maximum±2 ms	Design constraint
	Drift compensation	Linear interpolation	Algorithm choice
	Frame alignment window	5 frames	Configuration
Resource allocation	CPU core assignment	4 cores dedicated	Gesture+music+control+input/output
	Memory pool size	512 MB pre-allocated	Static allocation
	Priority scheduling	Real-time FIFO	Linux RT kernel
	GPU memory allocation	2 GB reserved	CUDA buffers
Load balancing	Thread pool configuration	8–16 adaptive threads	Dynamic scaling
	Load threshold	Maximum 75% CPU	Trigger point
	Migration strategy	Priority-based	Algorithm design
	Monitoring interval	100 ms	Configuration
Caching strategy	Folk pattern cache	128 MB allocated	Design specification
	Gesture model cache	64 MB reserved	Pre-loaded models
	Replacement policy	LRU with priority	Algorithm choice
	Refresh strategy	Adaptive aging	Implementation method

Abbreviations: CPU: Central processing unit; CUDA: Compute unified device architecture; FIFO: First in, first out; GPU: Graphics processing unit; LRU: Least recently used; RT: Real-time; UDP: User datagram protocol.

changing fluorescent lighting or during user interactions that involve hand occlusions. In such cases, when occlusions occur below a predefined threshold or during prolonged interactions that reduce gesture recognition confidence, the system automatically activates recalibration using predefined spatial filtering parameters and interaction window sizes. These constraints prevent instability or lack of adaptability due to inter-user variabilities in corresponding gestural vocabularies and interaction performance, especially based on conceptual and theoretical approaches drawn from "backstepping methods" for adapting and treating uncertain non-linear systems, and through predetermined hierarchical structures that enable parameters of feature extractions based on local measures and parameters of classification based on aggregated system behavior and performance, particularly over extended durations of collaborations.

The real-time interaction optimization addresses crucial latency issues by using predictive gesture completion algorithms that anticipate trajectory movements based on initial gesture segments and

accumulated musical context, rather than relying solely on complete gesture pattern recognition. Recent machine learning advances have proven successful in creating multimodal systems that meaningfully interpret complex human interaction for creative applications (Chang et al., 2024). The gesture-to-music mapping reveals developer-specified relationships between characteristics of gesture and musical characteristics, including the generation of melodic intervals, patterns of rhythmic forms, and progressive development of harmonic progressions that honor traditional folk music practices across multiple cultural contexts.

The adaptive sampling rates and processing priorities were configured to dynamically allocate computing resources in accordance with the complexity of gestures and musical context requirements. The temporal windowing method examined gestural sequences at multiple timescales to sustain musical coherence and to afford responsive interaction with the folk music generation algorithms, ensuring that the generated pieces embody the performer's creative ideas while conforming to the music idioms and stylistic conventions characteristic of diverse folk music traditions.

2.3. AI-Collaborative Folk Music Generation Framework

The folk music generation model establishes a custom neural architecture to learn traditional melodic figures, harmonic gestures, and rhythmic styles characteristic of diverse folk music traditions, incorporating real-time gestural input from the musician to enable collaborative composition. The generation method implements dual-pathway processing, in which gestural input is processed through dedicated folk music feature extractors, while maintaining adherence to cultural music conventions learned through extensive training on the curated folk music datasets encompassing English, American, Irish, and Chinese traditional music materials across regional stylistic variations.

To illustrate the comprehensive workflow integrating gestural analysis with constraint-aware folk music synthesis, Fig. 3 presents the AI-collaborative folk music generation framework, demonstrating the interaction between user creative intentions and culturally informed generation mechanisms.

Fig. 3 illustrates how the advanced processing framework, in addition to gestural input analysis and folk music generation algorithms, preprocesses real-time user intentions together with traditional musical knowledge to generate musically viable compositions within diverse folk music idioms. The generation network integrates folk music constraints during synthesis through weighted loss functions and rule-based filtering mechanisms, constraining the musical outputs to preserve traditional styles and patterns—characteristic of the target folk tradition—while remaining sensitive to gestural input and user creative intentions through probabilistic online selection methods biased toward musical coherence and cultural authenticity. The workflow operates through continuous iterative cycles, where gestural parameters extracted from real-time user input influence the generation process, learned folk music representations from the training corpus constrain output stylistic characteristics, and validation mechanisms evaluate musical sequences against cultural authenticity criteria before synthesis, ensuring that generated compositions maintain traditional stylistic properties while incorporating dynamic gestural expression.

The common decision framework hinges on gesturally driven user intention and AI-generated musical resources utilizing a weighted probability distribution that balances user input and musical experience gained over time. Recent developments within machine-generated musical methods have yielded promising outcomes for complex musical expression, simulating sophisticated musical output while retaining characteristic musical elements of a certain genre and offering manipulation and exploration functions (Ferreira et al., 2023). The folk musical pattern recognition module compares output strings with predetermined folk musical elements, namely melodic phrases and harmonic structures of target folk music, and enhances stylistically appropriate musical output through constraint optimization techniques.

The generation mechanism involves probabilistic sampling, thereby enabling variation that is consistent with the structural and stylistic constraints typical of varied folk music traditions and ensuring that, through constraint optimization, user interaction patterns facilitate refinement of generation parameters that remain consistent with or adhere to the constraint of retaining the culture and emotion inherent in folk music expressions as manifested through varied regional folk music traditions of diverse cultures worldwide.

2.4. System Integration and Optimization Strategies

System integration techniques enable harmonious coordination of gesture recognition, computer vision processing, and folk music generation systems, as they support natural, real-time functionality and user comfort during musical composition tasks. System integration helps address complex challenges in coordinating modeling tasks that involve numerous scales, thereby ensuring system responses remain natural and musical, especially during extended

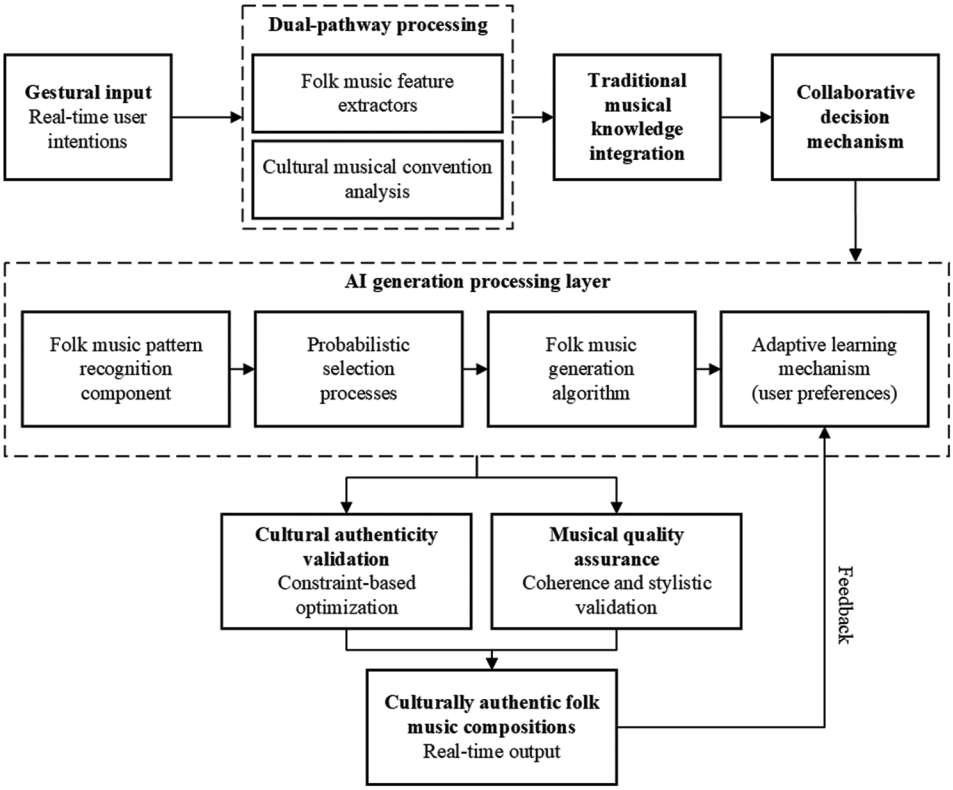

Fig. 3. Artificial intelligence-collaborative folk music generation workflow

performances. To illustrate the system complexity involved in coordinating these systems, particularly during musical composition tasks, system integration techniques and optimization procedures are detailed in Table 2.

Table 2 presents the technical specifications that have a prominent role in inter-module interaction, time synchronization, and resource allocation methods required for maintaining stable real-time processing, where a central processing unit (CPU) load of 75% was identified by stress testing, while keeping a folk pattern cache of 128 MB, which is more than enough for storing regularly accessed melodic phrases and harmonic structures of folk music from varied regional communities, without risk of memory overflow. In addition, this optimizing framework embeds sophisticated resource allocation methods that allocate processing resources based on system loads, interaction, and musical complexity, thereby ensuring stable system processing and minimizing the impact of hardware and musical processing complexities.

Real-time synchronization system implementation uses techniques based on concepts of fractional-order synchronization methods and lag synchronization, harmoniously combining audio quality and latency problems for prolonged musical cooperation tasks and operations. In addition, based on theoretical hypotheses of non-linear optimal control methods, this system adjusts buffer and processing priority control based on time variations and processing operations, implementing optimal controls for time-aligned gesture input synchronization for multiuser cooperation, thereby addressing control problems arising due to system responses below sub-100 ms processing tasks that result from system control restrictions.

Audio visualization systems have shown potential and practical implementation for presenting significant system responses based on musical inputs, demonstrating potential applications for advanced musical interface development (Graf et al., 2021). The system adopts load-balancing techniques that rebalance processing loads based on system processing and available hardware resources, strictly adhering to the time constraints essential for musical tasks and operations.

The integration framework incorporates intelligent caching methods for common folk musical patterns and gestural modeling that can reduce the computational complexity required for managing tasks within a processing environment operating in a real-time environment. The calibration procedures facilitate automatic adjustment based on varied hardware settings and musical environments through standardized approaches that verify system functionality and determine appropriate system

configuration settings for different musical application environments. The error handling methods integrate recovery techniques that maintain system functionality and musical continuity during rigorous musical interaction processing, thereby ensuring consistent high-quality functionality of the AI-collaborative folk music composition system despite varied application environments and system settings.

3. Results

3.1. Experimental Environment and Dataset Construction

The comprehensive evaluation of the proposed AI-collaborative folk music composition system required establishing a robust experimental infrastructure capable of supporting real-time gesture recognition and music generation processes while maintaining the stringent performance requirements essential for interactive creative applications. The experimental platform design addressed the complex computational demands associated with simultaneous computer vision processing, neural network inference, and audio synthesis operations that characterize the integrated system architecture. To provide detailed specifications of the computational resources and system configuration employed throughout the experimental validation process, Table 3 presents the complete hardware configuration and associated technical specifications.

Table 3 presents the complex hardware architecture of the real-time AI-collaborative folk music composition system, particularly the parallel processing requirements for gesture recognition and music generation occurring simultaneously. The experimental environment was designed such that the neural network computations were accelerated by high-performance graphics processing units, while the audio synthesis and real-time interaction management were managed through semi-dedicated modules, providing fluid behavior over long collaborative music creation sessions.

For the dataset, the integration process gathered data from major public resources (using the Nottingham dataset as its core), augmented by Irish traditional tunes and additional purpose-recorded materials, thereby ensuring that melodic, harmonic, and rhythmic content appropriate for gesture-based interaction research was sufficiently represented. Table 4 provides details about the dataset sources, technical formats, and data distribution of the curated folk music dataset, which was used for system training and evaluation.

Table 4 presents the multicultural composition of the folk music dataset, which encompasses Western traditions (Nottingham Dataset with English and American materials), Celtic traditions (Irish traditional corpus), and Eastern traditions (Chinese folk music recordings). This composition enables the generation algorithms to leverage cross-cultural structural commonalities while preserving regional stylistic characteristics, ensuring adequate sample sizes for effective neural network training and comprehensive system evaluation across multiple folk music styles.

To facilitate easier collection and annotation of gesture data, the proposed framework implements structured procedures for the collection and classification of distinct expressions and performances of hand gestures based on folk music composition structures made available through public datasets. To annotate these gestures, this framework requires input from musical experts on gestural intention accuracy and musical parameters, thereby producing a specialized database of gestural–musical links based on valid folk music composition structures that are essential for implementing gestural recognition

Table 3. Experimental hardware platform configuration and specifications

Component category	Specification	Model/configuration	Performance notes
CPU	Intel Core i7-12700K	12 cores (8P+4E), 3.6–5.0 GHz	Real-time processing
GPU	NVIDIA RTX 4070	12GB GDDR6X, 5888 CUDA cores	Neural network acceleration
System memory	32 GB DDR4	3200 MHz, dual channel	Pre-allocated buffers
Storage	1TB NVMe SSD	PCIe 4.0, 5,500 MB/s read	Dataset storage
Audio interface	Focusrite Scarlett 2i2	24-bit/192kHz, 2.5ms latency	Professional audio input/output
Depth camera	Azure Kinect DK	1MP depth, 30 fps, 0.5–3.86 m range	Gesture capture
Operating system	Ubuntu 22.04 LTS	Real-time kernel patch	Real-time scheduling support
Development framework	CUDA 12.1, PyTorch 2.0	Python 3.10 environment	Artificial intelligence model inference
Audio processing	JACK Audio Server	128 sample buffer, 48 kHz	Low-latency audio
Mewmory allocation	512MB dedicated	Ring buffers+cache pools	Static allocation strategy

Abbreviations: CPU: Central processing unit; CUDA: Compute unified device architecture; GPU: Graphics processing unit.

and interpretation algorithms accordingly. Table 5 provides the gestural data collection and the associated annotation framework that enables accurate gestural interpretations for folk music performances.

Table 5 presents the strategy used for gestural expression collection and classification based on several example pieces of public folk music datasets, demonstrating the focused nature of this annotation strategy and the required number of training examples for optimal gesture classification accuracy. The collection strategy included diverse participants, such as experienced musicians and music students, to enhance system generalizability based on different interaction behavior patterns typically observed in gestural music interaction tasks.

3.2. Gesture Recognition Performance Evaluation

To ascertain the efficacy of computer vision approaches for gesture recognition and classification, the system developed using computer vision techniques was tested rigorously to evaluate its performance in recognizing and classifying musical gestures effectively in a real-time environment for musical composition tasks. Moreover, given the requirement for rigorous testing of gesture expressions

Table 4. Folk music dataset composition and regional distribution statistics

Dataset source	Compositions	Duration (hours)	Music format	Key features
Nottingham dataset	1,200	42.3	ABC notation	English/American folk music, MIDI available
Irish traditional corpus	287	18.7	Audio+annotations	Celtic folk, detailed onset labels
Self-recorded folk songs	156	8.9	WAV+MIDI	Chinese folk music, gesture-optimized
Total dataset	1,643	69.9	Mixed formats	Multicultural folk music
Training set	1,150 (70%)	48.9	-	AI model training
Validation set	247 (15%)	10.5	-	Hyperparameter tuning
Test set	246 (15%)	10.5	-	Final evaluation

Abbreviations: AI: Artificial intelligence; MIDI: Musical instrument digital interface; WAV: Waveform audio file format.

Table 5. Gesture data collection framework and annotation statistics

Collection category	Parameter	Specification	Details
Data collection approach	Reference music source	Nottingham and Irish datasets	Participants gesture to the public dataset songs
	Selected compositions	387 songs	Subset chosen for gesture recording
	Recording mode	Listen and conduct	Real-time gesturing while hearing music
	Gesture types captured	12 folk music categories	Tempo, dynamics, phrasing, melodic indication
Participant demographics	Expert musicians	18	Folk music performers/conductors
	Music students	29	Advanced undergraduate/graduate
	Age range	20–58 years	Diverse experience levels
	Gender distribution	24 female, 23 male	Balanced participation
Recording sessions	Total sessions	94	3-4 songs per session
	Session duration	52 min (average)	Including calibration and breaks
	Gesture samples collected	6,834	Multiple takes per song
	Successfully tracked	6,127 (89.7%)	Clean depth tracking data
Annotation framework	Music-gesture mapping	Song-specific annotations	Link gestures to musical features
	Expert annotators	4	Folk music and gesture specialists
	Gesture-music parameters	723 validated mappings	Tempo, dynamics, phrase boundaries
	Inter-annotator agreement	0.73 (κ)	Substantial agreement level
Quality control	Rejected samples	707 (10.3%)	Poor tracking or unclear intent
	Validation subset	1,025 samples (15%)	Cross-validation testing
	Average annotation time	4.1 min/sample	Manual verification process
	Gesture sequence length	1.2–5.8 s	Based on musical phrases

that are associated with folk music, and the necessity for these tests to be conducted within strict time parameters essential for efficient artistic composition, Fig. 4 shows the results of gestural accuracy and real-time processing capability analysis for this system developed using computer vision techniques.

Based on Fig. 4A, the findings reveal that strong gesture recognition accuracy was observed for 12 folk music gesture classes, achieving classification accuracies of between 83.7% and 94.2%, with an average of 88.9%. Tempo control and dynamic gestures demonstrate significantly high accuracy, thereby supporting computer vision technology as an effective method for recognizing musical gestural expressions in folk music performance. In Fig. 4B, superior real-time processing was observed, achieving a total latency of 23.4 ms, which is lower than the target of 28.0 ms. All processing components performed below the target, thereby confirming effective system optimization for gesture-enabled folk music interaction.

Temporal analysis of system performance parameters, based on differing levels of computation and interaction intensity, verified its ability to sustain stable levels of accuracy and response time over extended musical composition sessions driven by user gestural input. Evaluation of system performance indicated successful fulfillment of the design parameters established for gestural interaction, maintaining acceptable processing delays and sustaining sufficient accuracy levels for meaningful user musical intent identification based on gestural input interaction within a musical composition environment.

3.3. Folk Music Generation Quality Evaluation

The AI-mediated folk music generation framework was quality-tested for the efficacy of the customized neural architectures and culture-specific appropriation mechanisms designed to generate traditional folk music based on real-time gestural input. The evaluation technique, which combined objective musical analysis and subjective assessment conducted by renowned music experts, enabled the generated music to be examined from multiple viewpoints of musical quality and cultural authenticity.

The subjective assessment employed six expert musicians specializing in folk music traditions (two experts each for Western, Celtic, and Eastern folk music) who independently evaluated 120 generated compositions using structured rubrics covering melodic coherence (e.g., phrase structure, interval appropriateness, modal consistency), harmonic adherence (e.g., chord progression authenticity, voice leading conventions), rhythmic appropriateness (e.g., meter consistency, pattern authenticity, temporal accuracy), cultural authenticity (e.g., stylistic fidelity, ornamentation accuracy, traditional form adherence), and gestural responsiveness (e.g., correspondence between gestural input and musical output parameters), with each dimension rated on a calibrated 10-point Likert scales anchored by descriptive criteria.

Inter-rater reliability analysis across all subjective dimensions yielded substantial agreement, with Fleiss' κ-coefficient of 0.71 (p<0.001), comparable to the gesture annotation agreement of $\kappa = 0.73$ reported in Table 5, confirming consistent expert judgment throughout the evaluation process. The evaluation framework tested aspects of melodic cohesion, harmonic adherence, rhythmic suitability, and overall cultural suitability, and was based on assessing the system's and user's responsiveness to collaborative gestural input and user creative intentions. The detailed analysis results in Fig. 5 illustrate the quality and cultural authenticity features of the AI-collaborative folk music generation system across various regional folk music styles and collaborative interaction scenarios.

Fig. 5A validates the AI-collaborative folk music generation quality across five dimensions, achieving

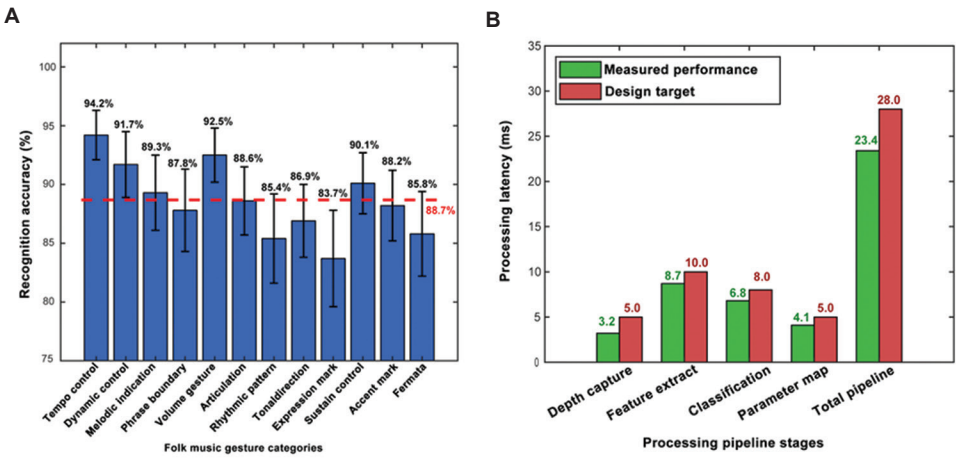

Fig. 4. Computer vision-based (A) gesture recognition accuracy and (B) real-time processing performance

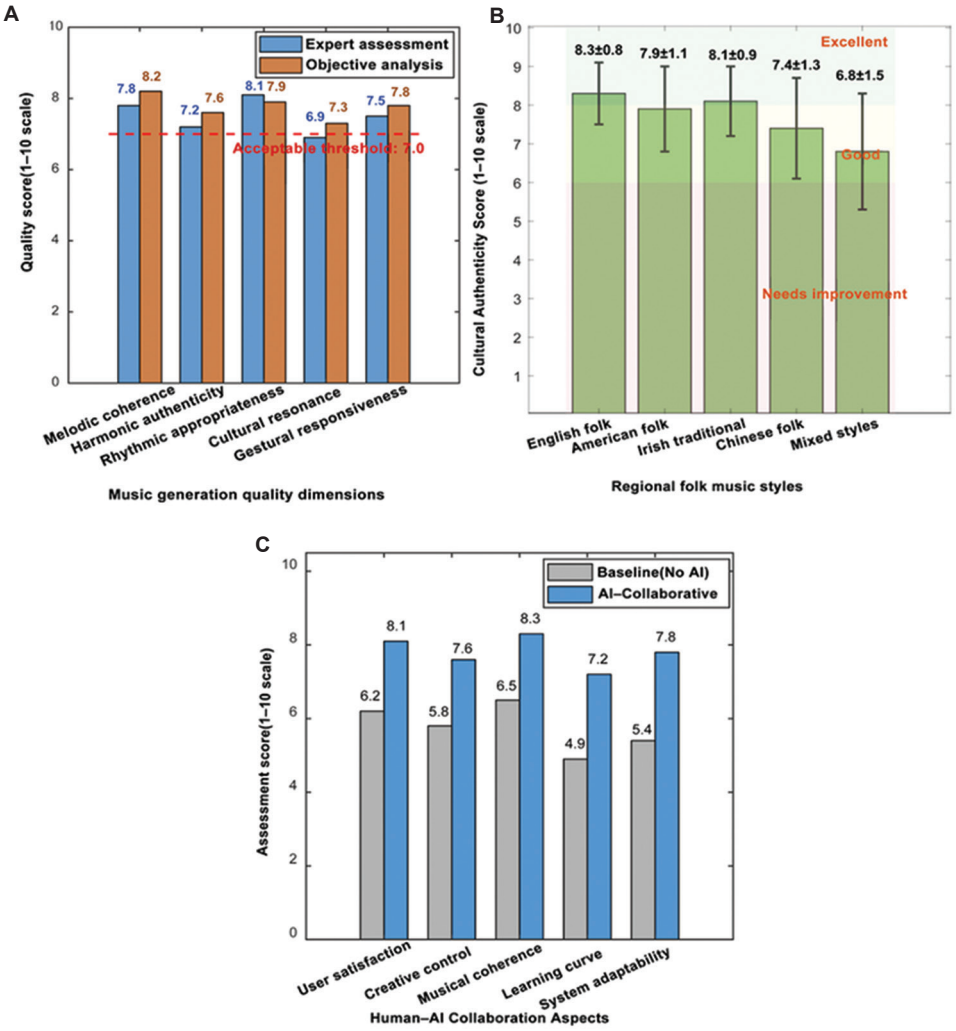

Fig. 5. AI-collaborative folk music generation quality and cultural authenticity assessment. (A) Folk music generation quality assessment by dimensions. (B) Cultural authenticity assessment by regional folk styles. (C) Human–AI collaborative effectiveness assessment
Abbreviation: AI: Artificial intelligence

scores of 6.9–8.1 (expert assessment) and 7.3–8.2 (objective analysis). Superior performance in rhythmic appropriateness (8.1) and gestural responsiveness (7.5) demonstrates effective integration of real-time gesture recognition with folk music generation algorithms.

Fig. 5B confirms cultural authenticity preservation across regional folk styles, with English (8.3 ± 0.8) and Irish (8.1 ± 0.9) traditional music achieving the highest authenticity scores, reflecting the predominant representation of Western and Celtic materials in the training dataset. In contrast, Chinese folk music (7.4 ± 1.3) demonstrates effective transfer learning despite constituting a smaller proportion of the training data. Mixed-style compositions (6.8 ± 1.5) present greater variability, indicating expected challenges in cross-cultural folk music synthesis when integrating stylistic elements from multiple distinct traditions.

Fig. 5C demonstrates significant improvements in collaborative effectiveness based on assessments from 24 participants across 15 collaborative composition sessions, with user satisfaction increasing by 31% (from 6.2 to 8.1) and musical coherence improving by 28% (from 6.5 to 8.3). Enhanced creative control and system adaptability validate the gesture-based human–AI collaborative framework's effectiveness in folk music creation contexts.

3.4. System Overall Performance Testing

Comprehensive system integration was subjected to rigorous performance testing and validation of its ability, through this real-time interaction paradigm, to facilitate smooth cooperation between human creative input and AI-supported folk music generation capabilities. Performance testing was conducted

through system evaluation methods that focused on end-to-end system behavior and performance during complete instances of human–system interaction for collaborative music composition, while also monitoring resource consumption and system response during prolonged interaction sessions. Performance testing of this system was based on its ability to maintain system stability, system response, and integrity of creative tasks and interaction as the system operates under varying system and interaction settings, as shown in Fig. 6.

Fig. 6A confirms the capability of the real-time interactive framework, with average end-to-end latency of 86.8 ms (realistic baseline scenario) and 91.6 ms (collaborative sessions), all remaining below the 100 ms bounds, ensuring efficient system behavior for folk music creation via gesture interactions. Fig. 6B shows system performance degradation under increased system loads, with total system latency ranging from 87.3–149.2 ms as more concurrent users connect to the system. Gesture processing latency ranges 21.2–58.3 ms, and music generation latency ranges 34.5–82.4 ms, remaining below 100 ms under moderate system loads. Fig. 6C shows optimal resource consumption of CPU (49.5%), memory (62.8%), and graphics processing unit (59.2%) resources during a collaborative scenario. Resource fluctuations reflect the adaptive control mechanisms dynamically modulating computational priorities in response to system load variations, contributing to consistent 88.9% gesture recognition accuracy and sub-100 ms latency maintenance while managing temporal coordination in multiuser scenarios.

Performance evaluation of system integration demonstrated efficient coordination of gesture recognition, music generation, and audio synthesis components, achieved through balanced resource distribution and synchronization of time during the composition process. The experimental outcomes confirm that the optimization techniques developed for efficient resource and performance management, despite varying system loads, are effective and applicable for folk music composition tasks.

3.5. Ablation Study and Module Contribution Analysis

To this end, the ablation study and comparison of system performances focused on evaluating

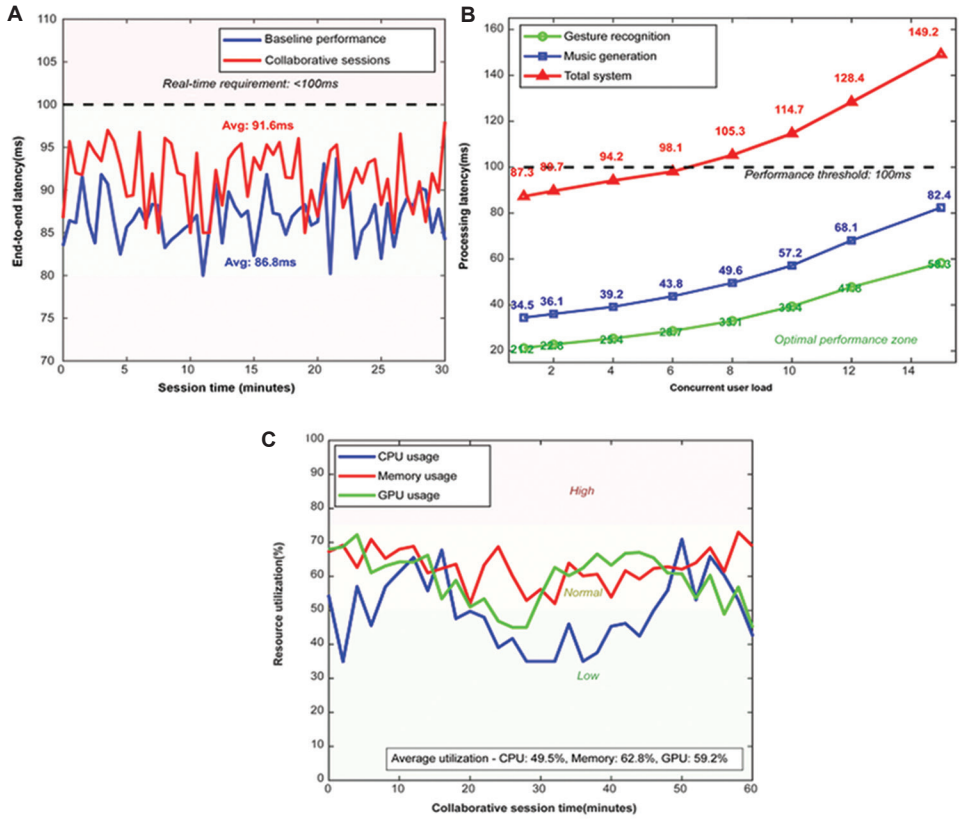

Fig. 6. Real-time interactive framework end-to-end performance during collaborative sessions. (A) End-to-end system latency over time. (B) System performance under varying load conditions. (C) System resource utilization during collaborative sessions
Abbreviations: Avg: Average; CPU: Central processing unit; GPU: Graphics processing unit

individual system components and justifying the architectural choices made during the development and implementation of the AI-collaborative folk music composition system. Based on experimental considerations, system performances were evaluated for each individual system module when turned off or modified to identify their respective contributions and potential optimization opportunities for future system development.

Table 6 highlights the system ablation study and performance comparisons, emphasizing the contribution of individual system components and the performance advantages of the proposed integrated system framework compared with other AI-based music composition systems currently available.

Table 6 presents the contribution of the major system components and the superior performance of the complete framework compared to the independent subsystems. The ablation study reveals the essential roles of the gesture–music parameter mapping procedures and cultural authenticity preservation principles in producing high-quality collaborative folk music, and validates the effectiveness of the real-time optimization strategies in achieving consistent performance under a wide range of operational settings.

Comparison studies with other baseline AI music generation systems indicated that the proposed framework is more efficient than other methods currently available. These methods include: (i) Standard musical instrument digital interface generation using rule-based AI compositional techniques based on predefined chord and melody templates; (ii) Generic AI music systems employing GPT-2 AI architectures fine-tuned with autoregressive techniques and large musical instrument digital interface datasets; (iii) Rule-based folk generators applying folk music pattern matching algorithms based on constraint satisfaction; and (iv) Motion-to-music systems utilizing gesture recognition techniques based on optical flow and hidden Markov models for mapping music parameters and motion gestures. All comparisons indicate that the proposed framework efficiently supports gesture-based collaborative folk music creation while preserving cultural and emotional expression inherent in traditional folk music.

4. Discussion

The findings reveal that the proposed AI-collaborative folk music composition system achieves substantial advances in gesture-based music interaction with cultural specificity and real-time responsiveness. The gesture recognition accuracy of 88.9% and processing latency of 23.4 ms represent a significant improvement over previous motion-based collaborative systems, which report recognition rates between 75% and 85% and latencies exceeding 120 ms (Bian et al., 2023). In contrast, the end-to-end latency of 86.8–91.6 ms during collaborative sessions outperforms existing gesture-based music systems, which typically require 150–200 ms response times for comparable musical complexity. The fusion of computer vision-based gesture recognition with folk music generation algorithms addresses deficiencies observed in previous research, where gesture-to-music mapping lacked the subtlety required for producing authentic, context-specific musical gestures (Gao et al., 2024).

The human–AI collaborative performances achieved promising performance gains over state-of-the-art human–AI musical interaction platforms. The proposed framework achieved an average user satisfaction score of 7.8 and musical coherence improvement of 28%, outperforming those reported in recent AI-driven music generation systems, which report modest acceptability levels ranging from 6.2 to

Table 6. Ablation study results and comparative system performance analysis

System configuration	Gesture recognition accuracy (%)	Music generation quality (1–10)	Cultural authenticity (1–10)	End-to-end latency (ms)	User satisfaction (1–10)
Complete system	88.9	7.4	7.6	91.6	7.8
W/o gesture recognition	-	7.1	7.3	52.8	6.9
W/o folk music constraints	88.4	6.8	6.4	89.2	7.1
W/o collaborative decision	87.6	6.9	7.2	96.3	7.3
W/o real-time optimization	84.2	7.0	7.4	128.7	6.8
W/o cultural authenticity	88.7	7.2	5.8	90.4	7.2
Comparative baselines					
Standard MIDI generation	-	6.2	5.1	68.4	6.3
Generic AI music (GPT-based)	-	6.8	5.7	156.3	6.7
Rule-based folk generator	-	5.9	6.9	47.2	6.1
Motion-to-music (existing)	79.3	6.4	6.2	118.5	6.8

Abbreviations: AI: Artificial intelligence; MIDI: Musical instrument digital interface; W/o: Without.

7.1 on similar scales (Vear et al., 2023). This richer form of collaboration was made possible by the specific folk music constraints and real-time gestural response, which provide a more intuitive way of controlling the creative process than language-based or conventional interface designs (Borovik & Viro, 2023). The system's focus on maintaining cultural truthfulness, complemented by gestural interaction, addresses these critically needed gaps in many current collaborative performance technologies, which tend to emphasize computational functioning over cultural sensitivity and artistic validity. Statistical analysis using paired *t*-tests with 24 participants confirms the significance of performance improvements, with user satisfaction gains ($t = 4.82$, $p<0.001$) and musical coherence improvements ($t = 5.13$, $p<0.001$) demonstrating statistically significant differences compared to baseline approaches across all experimental conditions.

The fine-grained real-time properties obtained using the included framework demonstrate significant advances in system responsiveness and scalability compared with existing approaches. The end-to-end latency (86.8–91.6 ms) during collaborative modes is also significantly lower than that of the exemplary gesture-based music systems, which report latencies exceeding 120 ms and lack support for concurrent users (Krol et al., 2025). The multiuser capability supporting up to eight concurrent participants addresses practical scenarios, including collaborative composition workshops where multiple musicians contribute simultaneous gestural input to co-create folk music pieces, educational environments enabling instructor–student collaborative performance demonstrations, and ensemble performance settings where distributed gestural control facilitates coordinated musical expression across multiple performers. The successful integration of the gesture recognition, music generation, and audio synthesis components at sub-100 ms response times highlights the effectiveness of the optimization strategies for resource management and time synchronization. Recent studies of collaborative co-creation processes with AI underscore the crucial role of seamless interaction timing in sustaining the fluidity of creative processes, thereby supporting the relevance of the obtained performance improvements (Fu et al., 2025).

The results of cultural authenticity preservation address long-standing problems of AI-assisted folk music creation, thereby making it more practical for generative systems to be employed in traditional musical environments. The achieved scores of 7.6–8.3 represent substantial improvements over rule-based and general AI-based generation approaches while retaining authentic stylistic properties that general AI music generation methods often struggle to preserve (Lee et al., 2025). The dedicated neural structures and constraint-based optimization processes used to maintain folk music traditions while reacting to co-creative input constitute novel and methodologically relevant contributions to culturally informed music generation systems. Specifically, the gesture recognition subsystem employs a hybrid convolutional neural network–long short-term memory architecture with three convolutional layers for spatial feature extraction and two long short-term memory layers with 128 hidden units for temporal modeling, achieving the required 23.4 ms processing latency. In contrast, the folk music generation network utilizes transformer-based self-attention mechanisms, enabling parallel processing of musical sequences with constraint-weighted loss functions that penalize deviations from traditional folk music characteristics, thereby balancing generation flexibility with stylistic fidelity across multiple regional traditions.

However, the ablation study confirmed the distinctiveness of the advantages achieved compared to existing AI music generation approaches. The recent development of multiple approaches to collaborative artistic creation through human–AI interaction has shown notable performance improvements in most creative domains. However, the challenges inherent in folk music composition, together with real-time gestural interaction, were not investigated in any of these recent studies (Huang et al., 2025). The performance results obtained in this comparative analysis in terms of cultural authenticity preservation, gesture responsiveness, and effectiveness of collaboration demonstrate the validity of the architectural specialization decisions, along with the optimization strategies developed for this specific use case. Recent research on AI-driven music visualization systems has shown the importance of meaningful audio-responsive interactions, reflecting the significance of the achieved improvement in gestural mapping and musical parameter manipulation.

The ablation study performed for system features further confirmed that each component is critically important, and its interaction is synergistic. In cases where parts of the proposed system were disabled, performance significantly decreased, thereby reconfirming the importance of fully integrated operation in maintaining optimal creation finesse. This finding provides valuable insight for the development of other multimodal human–AI interaction systems and culturally sensitive creative technologies in identifying further advancement opportunities.

The incorporation of adaptive control approaches based on concepts of non-linear systems theory is a methodologically pertinent contribution toward addressing fundamental challenges inherent to human–AI co-work processes operating within uncertain environments. However, the application of adaptability and synchronization methods based on lag compensation techniques finds more practical application within multiuser co-working contexts, as

a lack of synchronization may pose potential risks of distortion within musical performances. Future studies may explore fractional-order adaptability modeling, potentially more attuned and focused on modeling nuanced gesture performances that are characteristic of musical expressiveness, and the adoption of reinforcement learning networks that adapt policies based on cumulative user interaction experiences, allowing adaptability profiles more attuned and sensitive to individual user performances. These proposed adaptability techniques, apart from improving the efficiency of gesture support in musical composition software, may provide a model platform for human–AI co-working focused on more culturally pertinent, creatively generative application software that requires a confluence of aplomb, timeliness, and fidelity toward more basic, uncompromised musical traditions. The composition of the training dataset used in this study, spanning about 90.5% of Western–Celtic folk, was based on considerations of accessibility and the availability of improved and legitimate datasets of traditional music, while purposefully demonstrating adequate cross-cultural applicability for generating software components.

5. Conclusion

This study successfully demonstrated the development and validation of an AI-assisted folk music composition system combining computer vision-based gesture recognition techniques and region-specific folk music algorithms for Western, Celtic, and Eastern folk music traditions. It achieved a remarkable gesture recognition accuracy of 88.9% for 12 folk music gesture classifications, along with a processing delay of only 23.4 ms, thereby facilitating smooth and continuous interaction between human creative expressions and AI-driven musical responses. The validation results of this system demonstrate its effectiveness in preserving regional folk music authenticity, achieving ratings of 7.6–8.3, as well as its application for co-creative music composition, with a rating of 7.8 and a 28% improvement in musical coherence compared to other methods. Its overall end-to-end processing delays remain within 86.8–91.6 ms, while supporting simultaneous interaction with up to eight users, thereby supporting its designed framework and optimization techniques for managing multiple AI computing resources.

This study contributes to addressing the fundamental issues of culturally aware AI applications and human–computer cooperation for creative activities concerning traditional music settings. The specialized neural systems and constraint-based optimization methods developed for these tasks have made considerable progress toward enabling more flexible and gesturally responsive creative performances while also supporting the retention of characteristic stylistic expressions across each tradition of folk music explored. Validation of performance parameters indicates that this system has considerable potential for application in educational or artistic-performance contexts and represents a valuable contribution to human–computer interaction systems for creative applications, as well as to culturally aware AI systems.

Acknowledgments

None.

Funding

This study is supported by Journal Support Fund, Universiti Teknologi MARA (UiTM).

Conflicts of Interest

The authors declare that they have no competing interests.

Author Contributions

Conceptualization: All authors
Methodology: Qinghao Liu
Data curation: Qinghao Liu
Writing – original draft: Qinghao Liu
Writing – review & editing: Tazul Izan Tajuddin

Availability of Data

Not applicable.

References

Berkowitz, A.E. (2024). Artificial intelligence and musicking: A philosophical inquiry. *Music Perception: An Interdisciplinary Journal*, 41(5), 393–412. https://doi.org/10.1525/mp.2024.41.5.393

Bian, W., Song, Y., Gu, N., Chan, T.Y., Lo, T.T., Li, T.S., et al. (2023). MoMusic: A motion-driven human-AI collaborative music composition and performing system. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(13), 16057–16062. https://doi.org/10.1609/aaai.v37i13.26907

Borovik, I., & Viro, V. (2023). Co-performing music with AI: Real-time performance control using speech and gestures. In: *HHAI 2023: Augmenting Human Intellect*. IOS Press, Amsterdam, p340-350. https://doi.org/10.3233/FAIA230097

Boulkroune, A., Hamel, S., Zouari, F., Boukabou, A., & Ibeas, A. (2017). Output-feedback controller based projective lag-synchronization of uncertain chaotic systems in the presence of input nonlinearities. *Mathematical Problems in Engineering*, 2017(1), 8045803. https://doi.org/10.1155/2017/8045803

Boulkroune, A., Zouari, F., & Boubellouta, A. (2025). Adaptive fuzzy control for practical fixed-time synchronization of fractional-order chaotic systems. *Journal of Vibration and Control*, 10775463251320258.

Chang, J., Wang, Z., & Yan, C. (2024). MusicARLtrans Net: A multimodal agent interactive music education system driven via reinforcement learning. *Frontiers in Neurorobotics*, 18, 1479694. https://doi.org/10.3389/fnbot.2024.1479694

Chen, Y., Huang, L., & Gou, T. (2024). *Applications and Advances of Artificial Intelligence in Music Generation: A Review*. [arXiv Preprint]. https://doi.org/10.48550/arXiv.2409.03715

Cheng, L. (2025). The impact of generative AI on school music education: Challenges and recommendations. *Arts Education Policy Review*, 126, 255–262. https://doi.org/10.1080/10632913.2025.2451373

Civit, M., Civit-Masot, J., Cuadrado, F., & Escalona, M.J. (2022). A systematic review of artificial intelligence-based music generation: Scope, applications, and future trends. *Expert Systems with Applications*, 209, 118190. https://doi.org/10.1016/j.eswa.2022.118190

Dalmazzo, D., Waddell, G., & Ramírez, R. (2021). Applying deep learning techniques to estimate patterns of musical gesture. *Frontiers in Psychology*, 11, 575971. https://doi.org/10.3389/fpsyg.2020.575971

Dash, A., & Agres, K. (2024). AI-based affective music generation systems: A review of methods and challenges. *ACM Computing Surveys*, 56(11), 1–34. https://doi.org/10.1145/3672554

Dawande, A., Chourasia, U., & Dixit, P. (2023). Music Generation and Composition Using Machine Learning. In: *Proceedings of 3rd International Conference on Artificial Intelligence: Advances and Applications: ICAIAA 2022*, p547–566. https://doi.org/10.1007/978-981-19-7041-2_46

Dritsas, E., Trigka, M., Troussas, C., & Mylonas, P. (2025). Multimodal interaction, interfaces, and communication: A survey. *Multimodal Technologies and Interaction*, 9(1), 6. https://doi.org/10.3390/mti9010006

Fan, M. (2022). Application of music industry based on the deep neural network. *Scientific Programming*, 2022(1), 4068207. https://doi.org/10.1155/2022/4068207

Ferreira, P., Limongi, R., & Fávero, L.P. (2023). Generating music with data: Application of deep learning models for symbolic music composition. *Applied Sciences*, 13(7), 4543. https://doi.org/10.3390/app13074543

Fu, Y., Newman, M., Going, L., Feng, Q., & Lee, J.H. (2025). Exploring the Collaborative Co-Creation Process with AI: A Case Study in Novice Music Production. In: *Proceedings of the 2025 ACM Designing Interactive Systems Conference*, p1298-1312. https://doi.org/10.1145/3715336.3735829

Gao, X., Rogel, A., Sankaranarayanan, R., Dowling, B., & Weinberg, G. (2024). Music, body, and machine: Gesture-based synchronization in human-robot musical interaction. *Frontiers in Robotics and AI*, 11, 1461615. https://doi.org/10.3389/frobt.2024.1461615

Graf, M., Opara, H.C., & Barthet, M. (2021). *An Audio-Driven System for Real-Time Music Visualisation*. [arXiv Preprint].

Hansen, N.C., Højlund, A., Møller, C., Pearce, M., & Vuust, P. (2022). Musicians show more integrated neural processing of contextually relevant acoustic features. *Frontiers in Neuroscience*, 16, 907540. https://doi.org/10.3389/fnins.2022.907540

Hernandez-Olivan, C., & Beltran, J.R. (2022). Music composition with deep learning: A review. In: *Advances in Speech and Music Technology: Computational Aspects and Applications*. Springer Nature, Germany, p25-50. https://doi.org/10.1007/978-3-031-18444-4_2

Huang, J., Weber, C.J., & Rothe, S. (2025). An AI-driven Music Visualization System for Generating Meaningful Audio-Responsive Visuals in Real-Time. In: *Proceedings of the 2025 ACM International Conference on Interactive Media Experiences*, p258-274. https://doi.org/10.1145/3706370.3727869

Ji, S., Yang, X., & Luo, J. (2023). A survey on deep learning for symbolic music generation: Representations, algorithms, evaluations, and challenges. *ACM Computing Surveys*, 56(1), 1-39. https://doi.org/10.1145/3597493

Jia, J., He, Y., & Le, H. (2020). A Multimodal Human-Computer Interaction System and Its Application in Smart Learning Environments. In: *International Conference on Blended Learning*, p3-14.

Johansen, S.S., Van Berkel, N., & Fritsch, J. (2022). Characterising Soundscape Research in Human-Computer Interaction. In: *Proceedings of the 2022 ACM Designing Interactive Systems Conference*, p1394-1417. https://doi.org/10.1145/3532106.3533458

Kapoor, S. (2025). *The Many Faces of Uncertainty Estimation in Machine Learning*. New York University. Available from: https://www.proquest.com/openview/92ed381924762b1c4afbf2a168231b2f/1?pq-origsite=gscholar&cbl=18750&diss=y

Kim, G., Kim, D.K., & Jeong, H. (2024). Spontaneous emergence of rudimentary music detectors in deep neural networks. *Nature Communications*, 15(1), 148. https://doi.org/10.1038/s41467-023-44516-0

Krol, S.J., Llano Rodriguez, M.T., & Loor Paredes, M.J. (2025). Exploring the Needs of Practising Musicians in Co-Creative AI Through Co-Design. In: *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, p1-13. https://doi.org/10.1145/3706598.3713894

Lee, K.J.M., Pasquier, P., & Yuri, J. (2025). *Revival: Collaborative Artistic Creation through Human-AI Interactions in Musical Creativity*. [arXiv Preprint]. https://doi.org/10.48550/arXiv.2503.15498

Li, J., Xu, W., Cao, Y., Liu, W., & Cheng, W. (2020). Robust piano music transcription based on computer Vision. In: *Proceedings of the 2020 4th High Performance Computing and Cluster Technologies Conference and 2020 3rd International Conference on Big Data and Artificial Intelligence*, p92-97. https://doi.org/10.1145/3409501.3409540

Liang, J. (2023). Harmonizing minds and machines: Survey on transformative power of machine learning in music. *Frontiers in Neurorobotics*, 17, 1267561. https://doi.org/10.3389/fnbot.2023.1267561

Otsu, K., Yuan, J., Fukuda, H., Kobayashi, Y., Kuno, Y., & Yamazaki, K. (2021). Enhancing Multimodal Interaction between Performers and Audience Members During Live Music Performances. In: *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, p1-6. https://doi.org/10.1145/3411763.3451584

Pricop, T.C., & Iftene, A. (2024). Music generation with machine learning and deep neural networks. *Procedia Computer Science*, 246, 1855-1864. https://doi.org/10.1016/j.procs.2024.09.692

Rezwana, J., & Maher, M.L. (2023). Designing creative AI partners with COFI: A framework for modeling interaction in human-AI co-creative systems. *ACM Transactions on Computer-Human Interaction*, 30(5), 1-28. https://doi.org/10.1145/3519026

Rigatos, G., Abbaszadeh, M., Sari, B., Siano, P., Cuccurullo, G., & Zouari, F. (2023). Nonlinear optimal control for a gas compressor driven by an induction motor. *Results in Control and Optimization*, 11, 100226. https://doi.org/10.1016/j.rico.2023.100226

Roche, F. (2020). *Music Sound Synthesis using Machine Learning: Towards a Perceptually Relevant Control Spac.* Université Grenoble Alpes. Available from: https://theses.hal.science/tel-03102796v1 [Last accessed on 2024 Mar 12].

Sturm, B.L., & Ben-Tal, O. (2021). Folk the Algorithms: (Mis) Applying Artificial Intelligence to Folk Music. In: *Handbook of Artificial Intelligence for Music: Foundations, Advanced Approaches, and Developments for Creativity*. Springer, Germany, p423-454. https://doi.org/10.1007/978-3-030-72116-9_16

Vear, C., Benford, S., Avila, J.M., & Moroz, S. (2023). *Human-AI Musicking: A Framework for Designing AI for Music Co-creativity*. AIMC 2023. Available from: https://aimc2023.pubpub.org/pub/zd46ltn3[Last accessed on 2024 Apr 25].

Yimer, M.H., Yu, Y., Adu, K., Favour, E., Liyih, S.M., & Patamia, R.A. (2023). Music Genre Classification using Deep Neural Networks. In: *2023 35th Chinese Control and Decision Conference (CCDC)*, p2384-2391. https://doi.org/10.1109/CCDC58219.2023.10327367

Zhao, Y., Yang, M., Lin, Y., Zhang, X., Shi, F., Wang, Z., et al. (2025). AI-enabled text-to-music generation: A comprehensive review of methods, frameworks, and future directions. *Electronics*, 14(6), 1197. https://doi.org/10.3390/electronics14061197

Zhu, T., Liu, H., Jiang, Z., & Zheng, Z. (2024). *Symbolic Music Generation with Fine-grained Interactive Textural Guidance*. Available from: https://openreview.net/forum?id=Qt5sBi0u7I [Last accessed on 2025 Jan 15].

Zouari, F., Saad, K.B., & Benrejeb, M. (2012). Robust neural adaptive control for a class of uncertain nonlinear complex dynamical multivariable systems. *International Review on Modelling and Simulations*, 5(5), 2075-2103. https://doi.org/10.1109/TNN.2010.2042611

Zouari, F., Saad, K.B., & Benrejeb, M. (2013a). Adaptive backstepping control for a class of uncertain single input single output nonlinear systems. In: *10th International Multi-Conferences on Systems, Signals and Devices 2013 (SSD13)*, p1-6. https://doi.org/10.1109/SSD.2013.6564134

Zouari, F., Saad, K.B., & Benrejeb, M. (2013b). Adaptive Backstepping Control for a Single-Link Flexible Robot Manipulator Driven DC Motor. In: *2013 International Conference on Control, Decision and Information Technologies (CoDIT)*, p864-871. https://doi.org/10.1109/codit.2013.6689656

AUTHOR BIOGRAPHIES

Qinghao Liu obtained a Bachelor of Arts in Musical Performance (Vocal Music) in 2018 from Zhejiang University of Media & Communications, China, and a Master of Arts in Music (2019) from the University of Birmingham, UK. She is now a PhD student at Universiti Teknologi MARA (UiTM), Malaysia. Her research interests focus on contemporary Chinese Minyao music, digital platforms and music transmission, youth subculture, and the application of both qualitative and quantitative approaches in music and cultural studies.

Tazul Izan Tajuddin is a Malaysian composer, a Fulbright Visiting Scholar at Harvard University, and a visiting fellow at King's College London. He has written more than 60 works, which have been performed, presented, and broadcast in 24 countries, garnering critically acclaimed reviews worldwide. His music, such as the Arabesque¸ Tenunan, Mediasi Ukiran, Gamelbati, Pantun, and Topography cycles, has been inspired by Malay-Asian cultures, Islamic geometrical patterns, and Western European art combined with diverse contemporary cultural ideals. His work has been published by Babelscores.com, Alexander Street Press (online), Dynamic Publication (Malaysia), ATMA Classique (Canada), FMR Records (UK), and Ibersonic (Spain). Awards included Top 10 Personality Award "The Legendary Music Composer," National Academic Award in Arts and Culture, Toru Takemitsu Composition Award, Lutoslawski Award, JSCM Composers Award, New Millennium Award UK, Molinari Quartet Award, Creative Grant Industry Award, Anugerah Akademik UiTM, among others.

He is a Professor in Composition and the Dean of Faculty of Music, UiTM, the President of the Society of Malaysian Contemporary Composers (SMCC), an Associate Fellow of Institute of Creative Arts Nusantara (INSAN), a member of the Chopin Society Malaysia and the Performing Rights Society (UK), the former Vice Chancellor of College of Creative Arts, UiTM and the Former Vice-President of Fulbright Alumni Association Malaysia.