

Review

Artificial Intelligence for Cross-Cultural Percussion Analysis: A PRISMA-Based Systematic Review of Tabla and Conga

Yuyang Chen , Veera Phansue*  and Pongsapich Kaewkulthorn 

Faculty of Fine Arts, Srinakharinwirot University, Bangkok 10110, Thailand

* Correspondence: veerap@g.swu.ac.th

Received: 19 March 2026; Revised: 8 April 2026; Accepted: 20 April 2026; Published: 8 October 2026

Abstract: While AI has enhanced automatic drum transcription, culturally unique percussion instruments, due to their common peculiar timbral, harmonic and microtiming properties, are still underrepresented. This systematic review explores the analysis of Hindustani tabla and Afro-Cuban conga (and other related non-Western percussion traditions) using an AI approach. The peer-reviewed papers retrieved from IEEE Xplore, ACM Digital Library, Scopus, ISMIR and Google Scholar according to PRISMA 2020 guidelines. Forty-seven studies were synthesized across the areas of transcription, rhythmic analysis, low resource learning, and generative modelling of adequate quality and the author applied certain preceding eligibility criteria. Convolutional recurrent neural networks (CRNNs) were able to distinguish between tableau and tabla-strokes in concert recordings with an F score of 0.965, while the other deep-learning models got an accuracy of 97.50% in rhythmic-pattern recognition. The groove modelling of Afro-Latin showed that the adaptability under limited annotated data was boosted by meta-learning and data augmentation, while tatum-level beat and microtiming analysis was critical. The generation of stylistic percussion patterns is another major challenge in the generative approach, which is also faced by sparse datasets, domain shift, cultural bias and real time polyphonic transcription. Therefore, for accurate and authentic cross-cultural percussion analysis culturally informed, transparent and human-centred models are crucial for AI.

Keywords: Artificial Intelligence; Music Information Retrieval; Cross-Cultural Percussion; Automatic Drum Transcription; Deep Learning; Meta-Learning; Micro-Timing Analysis

1. Introduction

The field of music information retrieval has been fundamentally reinvented by artificial intelligence and machine learning, which allow the extraction of high-quality representations of features with a wide range of applications in music pedagogy and production [1]. Although automatic drum transcription has been extensively studied in the case of Western popular music, the computational analysis of cross-cultural percussion instruments, such as the Hindustani tabla and Afro-Cuban conga, is an ongoing research field [2–4]. These instruments also have characteristic physical and timbral qualities that can be discerned, in contrast to the traditional Western drums of that era, and their rhythmic and acoustic complexities reward more advanced modelling approaches [5,6].

The Hindustani tabla provides a distinctly complicated acoustic environment owing to its mixed harmonic and percussion timbres, which allows musicians to emphasize particular places in a rhythmic cycle by playing different strokes differently [5]. One major issue with automatic stroke classification is that the harmonics of previous strokes may overlap and be erroneously classified in the case of a polyphonic concert recording [7]. In addition, the rhythmic pattern of Afro-Cuban culture is characterized by systematic deviations in microtiming that are not in a

strict metrical system in the West [7]. These digressions are not simple distortions but bear structural value in the stylistic constitution of genres such as Cuban guaguanco [8].

Other obstacles to having effective AI systems for such tools are the critical lack of data and the so-called domain drift phenomenon when such models are applied, which occurs when such models are trained on culturally different musical data [9–11]. To address these issues, scientists have proposed new techniques, such as model-agnostic meta-learning for fast adaptation in low-resource settings and generative MIDI models, such as ClaveNet, which uses data augmentation to augment small datasets [6,12,13]. The transparency of music-generative systems is also critical for ethical and culturally appropriate AI development [14,15]. This review adopts a systematic PRISMA-based approach to justify the contemporary research field between January 2020 and February 2026 and examines the role of AI models in meeting the specific challenges related to tablet and conga [16,17].

To mitigate the perceived gaps in the study of cross-cultural perceptions, the current study outlines the following research questions:

RQ1: How do modern artificial-intelligence models represent and control the timbral and harmonic characteristics of cross-cultural percussion instruments, such as the tabla and conga, compared with traditional Western drum kit models?

RQ2: To what extent can meta-learning and data augmentation strategies help address the lack of large annotated cross-cultural percussion datasets in cross-cultural percussion research?

RQ3: What representations and determinations of micro-timing and non-grid-based rhythmical structures characterize Afro-Latin musical cultures using extant computational methodologies?

This study adopts a thematic synthesis approach consistent with the systematic review methodology. The analysis is organized around three key dimensions:

1. **Timbral Representation and Stroke Modeling:** Exploring the ways in which computational models reproduce the hybrid harmonic-percussive nature of instruments such as the tabla.
2. **Could low-resource learning and model adaptability—the strategies of meta-learning, transfer learning, and data augmentation—be used to solve the problem of dataset scarcity?**
3. **Rhythmic Structure and Microtiming Analysis:** Exploration of computational models of non-metrical rhythmic systems, especially in Afro-Latin music cultures.

These dimensions inform the synthesis and interpretation of findings across the reviewed studies.

Moreover, this study specifically considers the issue of cross-cultural transferability in AI-percussion analysis and emphasizes the constraints of Western-centric models and the necessity of culturally agile computational models.

Based on the research objectives, this study makes the following key contributions:

- This study provides an overall global mapping of AI-driven research trends in non-Western percussion instruments, such as the tabla and conga, between 2020 and 2026.
- This indicates that current automatic drum transcription (ADT) systems have essential technical constraints, notably in highlighting harmonic overlap and maintaining culturally pertinent microtiming frameworks.
- This study provides interdisciplinary meta-learning and data augmentation for ethnomusicological audio datasets.
- This presents a hypothetical concept of the appraisal system, in which the transparency of music-generative AI systems and their cultural compatibility with various global percussion traditions are observed.

Key differences between Western and cross-cultural percussion systems are presented in **Table 1**.

Table 1. Comparison of AI Challenges in Western and Cross-Cultural Percussion.

Characteristic	Western Percussion	Cross-Cultural (Tabla/Conga)
Timbral Focus	Primarily percussive/inharmonic [18]	Harmonic and percussive overlap [5]
Rhythmic Organization	Rigid metrical grid [19]	Structural microtiming deviations [20]
Data Availability	High availability [2]	Critically scarce/Low-resource [21]
Modeling Priority	Frame-level classification [17]	Meta-learning and pattern-matching [6]

The current study presents a culturally mindful computationalism that synthesizes microtiming analysis, har-

monic resonance modelling, and AI-based transcription to produce symbolic representations based on the cultural context, namely, tala and clave patterns (see **Figure 1**).

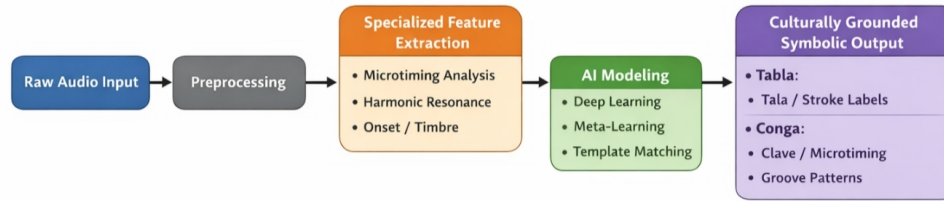


Figure 1. Conceptual Workflow for a Culturally Aware AI for Percussion.

2. Literature Review

The computational analysis of percussion has evolved from primitive methods of identifying onsets to advanced transcription systems that facilitate the recognition of subtle playing techniques in a broad range of musical practices. Shete and Deshmukh [22] used the zero-crossing rate and mel-frequency cepstral coefficients to extract audio representations relevant to tabla strokes. According to Vinod and Colaco [23], the accuracy of these methodologies is 93.33% for the prediction of *tāla* given the identified basic strokes. Although initial research was mainly focused on Western drum kits, modern researchers have expanded this to address the unique timbral and rhythmic complexity of other instruments, including the Hindustani tabla and Afro-Cuban conga.

2.1. Automatic Drum Transcription and Computational Stroke Identification

Non-negative matrix factorization and crude neural network methods have been the main methods used in traditional automatic drum transcription (ADT) systems to identify impulsive, inharmonic acoustic events, such as the bass drum and snare. The structure-conscious method of transcription suggested by Ishizuka et al. [19] applies a self-attention architecture to determine the quantized location of the Tatum on a grid of sixteenth notes. Cahyaningtyas et al. [2] built an ADT application using a segment-and-classify pipeline, obtaining an F1 score of 82.87 % on a variety of datasets. Rohit and Rao [5] introduced acoustic properties that can successfully model the harmonic and percussive peculiarities of Tabla hits, and they mentioned that pitch-shifting augmentation demonstrates hope in categorizing instruments without depending on instrument-specific data. Kaur et al. [24] use the tree-based algorithms to assign the strokes their distinct timbral signature. According to Riley and Dixon [25], there is still a significant difference between ADT accuracy and that of the pitch instrument, which can be explained by data constraints, including small samples and a lack of diversity. Moysis et al. [26] are interested in producing real-world data simulations to satisfy the scalable demands of modern state-of-the-art ADT models. Zehren et al. [27] stated that high-quality data acquired through crowdsourcing are essential for attaining reproducible transcription outcomes. Weber et al. [3] presented the standard test audio recording dataset (STAR Drums), offering a more detailed baseline for future transcription models.

2.2. Rhythmic Structure, Microtiming, and Metric Complexity

In Afro-Latin and Afro-Cuban traditions, rhythmic identity is defined by systematic microtiming differences that do not fit a strict grid of Western metrical rhythm. Hajek [28] performed a quantitative analysis of the temporal structures of Cuban guaguanco to describe the complexities in tempo tracking. Buisson et al. [29] developed a multilevel analysis approach for sections, downbeats, and beats in Latin American music. An example of such a combination, using ethnographic interviews and data-driven representations to dismantle Western metric assumptions, was presented by McGraw and Benadon [30], who stated that practitioners focus on the necessity to fill spaces rather than the beats of a clock. Burloiu [31] presented an expressive drum machine that varied its microtiming relative to a human musician and, thus, swung straight rhythms. Undy et al. [32] investigated the way in which music based on clave depends on the rhythmic patterns that are not grid-based to determine its stylistic character. The SAMBASET dataset proposed by Maia et al. [33] is a key source for the historical analysis of Samba de Enredo records. Danielsen et al. [34] prove that core acoustic variables like duration, attack also have a greater influence on the perception of microrhythm and grooviness of a rhythm. Mnasri et al. [35] explored the timeline-based ori-

entation of the playing of lead conga players as opposed to rigid measures.

2.3. Low-Resource Learning and Meta-Learning Strategies

A challenge facing cross-cultural music information retrieval (MIR) is the lack of large, high-fidelity annotated corpora. Grollmisch [6] introduces a meta-learning system, which is adapted to low-resource Tabla stroke transcription and tala recognition, thus enabling quick adaptation to other tasks. In another study, Kodag and Arora [21] show that Model-Agnostic Meta-Learning is capable of overcoming the problem of label heterogeneity and polyphonic auditory cases. Raimon et al. [1] provide a comprehensive survey that breaks down the mechanisms of generalization through meta-learning to a variety of audio-based tasks that would be performed by ontogenic learning processes using a minimum amount of data. Shi et al. [36] noted that meta-learning models have been shown to significantly outperform traditional supervised baselines in acoustic event detection. Maia et al. [9] suggest that culturally specific datasets should be selectively annotated based on the representation of the dataset, and allows beat-tracking algorithms to be trained on limited budgets. Samuel et al. [13] use meta-learners to improve performance of music source separation in low resource. Pinto [37] explored transfer learning in combination with human-in-the-loop detection of onset and used it in maracatu percussion to alleviate the workloads of annotation tasks.

2.4. Data Augmentation and Generative Rhythmic Models

The growing use of generative artificial intelligence (AI) models to produce adaptive accompaniments and generate authentic percussion patterns has been reported in various recent studies. To create Afro-Cuban-inspired MIDI drum grooves, García et al. [12] scaled the monotonic ideas of the groove transformer architecture and used a new data augmentation protocol to enhance the robustness of the model. Barnabò et al. [38] presented Cycle-DRUMS a framework, which is based on CycleGAN to automatically order the drum parts in a way that is plausible to a human sound and compliant with a given bass line. A systematic review of the field and practical uses of AI-generated music presented by Civit et al. [39] illustrates a visible change in the nature of the collaborative creative environment in the direction of becoming more autonomous. Jajoria and McDermott [40] examined text-conditioned symbolic drumbeat generation through latent diffusion models, which showed that rhythmic patterns generated by latent diffusion models can conform to given [textual] prompts. Oludare [41] utilized a generative style to analyze the timeline of the konkolo in Yoruba musical culture as a reference area for social communication cases. Morais et al. [42] augmented meter-tracking models to capture underrepresented time signatures in different music traditions, reducing domain drift and increasing cross-genre generalizability.

2.5. Cultural Diversity and Transparency in Music AI

The integration of artificial intelligence into global musical traditions requires a conscious focus on cultural congruence and ethical openness. Huang et al. [43] argue that the focus of the music information retrieval (MIR) academia should be refocused to responsible interdisciplinarity even as it strongly scrutinizes its epistemological assumptions. Born [44] confronts the outlines of computational music recognition based on the immanent principles of the musical blossoming of human beings, as opposed to economic growth. Duan et al. [10] introduced a special issue on cultural diversity, which also stated that musical cues cannot be interpreted fully in the absence of the associated cultural context. Porcaro et al. [45] show that artificial-intelligence systems are at risk of reinforcement of hegemonic biases, unless they are designed and assessed on the basis of a multicultural perspective of analysis. Battle-Roca et al. [15] conducted a systematic review of music-generative AI transparency to understand ethical and intellectual property issues. Ravikumar [14] proposed the use of DAIRHuM, a platform designed to reflect human musical judgments in a musical tradition, such as Carnatic music, using AI representations directly. Bai [46] also makes use of multimodal deep-learning algorithms to analyze perception of musical rhythm, and highlights how pattern-recognition algorithms should replicate the human mind.

3. Methodology

The present study broadly adheres to the guidelines of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (2020). Four stages comprise the review process: identification, screening, eligibility assess-

ment, and inclusion. A predetermined and systematic protocol was used to provide transparency, reproducibility, and methodological rigor. Civit et al. [39] emphasize the need of the so-called organized search approach to monitor the accelerated changes in the field of artificial intelligence in the world of music generation in 2020–2026. Smith [18] suggested that automatic drum transcription is a task that requires a specific approach to methodology that is capable of showing the variety of percussion cues that are not limited to traditional Western drum kits. As Muljono et al. [16] stressed, the classification of tempo using audio cues must rely on custom deep-learning architectures that are sufficiently versatile to cover as many rhythmic situations as possible.

3.1. Information Sources and Search Strategy

Grollmisch [6] defend the decision to use oriented scholarly databases to locate the works dealing with the field of low-resource percussion like the Hindustani Tabla. According to Rohit and Rao [5], the following search strings should be used: harmonic resonance to locate research on complex timbral stroke classification. Zehren et al. [27] concentrate on locating high-quality data collected by crowdsourcing to meet the requirements of transcription models used on a large scale. According to Batlle-Roca et al. [15], the sources of information should be general enough to determine the ethicality of music-generative AI systems. Weber et al. [3] chose STAR Drums data as a standard in the search for available literature related to automatic drum transcription. A diligent and reproducible search plan was created to find relevant studies. Searches were conducted in IEEE Xplore, ACM Digital Library, Scopus, ISMIR, and Google Scholar. The Boolean search query used was as follows:

("automatic drum transcription" OR "percussion analysis") AND ("tabla" OR "conga" OR "Afro-Cuban" OR "non-Western percussion") AND ("deep learning" OR "meta-learning" OR "music information retrieval")

The time frame of the search was January 2020 to February 2026. Exports of all retrieved records were subjected to duplicate removal and screening.

3.2. Eligibility Criteria

Riley and Dixon [25] show that the research using source separation as a method of extracting drum stems meets technical requirements needed by modern studies on transcription. Melucci et al. [47] describe the eligibility of the generative models which involve realistic synthetic data, which reduces the limitations of the small training sets. Kodag and Arora [21] argue that meta-learning studies are particularly qualified when overcoming the problem of label heterogeneity in percussion audio. The end-to-end comprehensive reviews mentioned by Raimon et al. [1] must include the inclusion criteria along with meta-learning, which is evaluated in the range of audio-processing tasks. Samuel et al. [13] provide the requirement that music source-separation models are only allowed to be qualitative when they demonstrate strong performance of low-resource conditions. Maia et al. [9] highlight the suitability of beat-tracking studies making use of selective annotations of restricted datasets.

Inclusion Criteria:

- Conference or journal publications (peer-reviewed).
- Research involving AI-based percussion analysis, transcription, or generation.
- Non-Western percussion instruments (e.g., tabla, conga, and Afro-Latin percussion) were included.
- Published between 2020 and 2026.

Exclusion Criteria:

- Research restricted to the analysis of Western drum kits.
- Non-AI-based approaches.
- Non-English publications.
- Lack of methodological description or assessment in studies.

3.3. Identification and Screening Process

Ravikumar [14] argues that the eligibility step should be used to qualify the similarities in the representations of machines compared to human musical judgments. Porcaro et al. [45] present the issue of using screening of titles and abstracts of articles on features of the artist character to reduce cultural biases concerning the electronic-

music research. Huang et al. [43] are in support of the interdisciplinary screening in order to identify the research that tackles the fractured realities of the global music representation. Born [44] appraised the quality of the identification stage with regard to the real-life problems that may arise with the diversification of music information retrieval. Duan et al. [10] provide screening criteria of studies which foster cultural diversity and contextual awareness in musical signals. A detailed PRISMA-based screening procedure was adopted to guarantee transparency and methodological rigor, as shown in **Figure 2**. Database searching identified 450 records, which were added to by 30 more records found elsewhere to make up the initial 480 records. All duplicates and initial filtering were conducted, and 420 records were left to be screened according to titles and abstracts. In this step, 326 records were eliminated as irrelevant. Subsequently, 110 full-text articles were evaluated based on eligibility, of which 63 articles were rejected on particular grounds, such as lack of relevance to the research topic ($n = 38$) and inadequate amount or methodology description ($n = 25$). Finally, 47 papers were found to fit all inclusion criteria and were added to the qualitative synthesis. Such a systematic approach will promote the reliability and reproducibility of the study selection process.

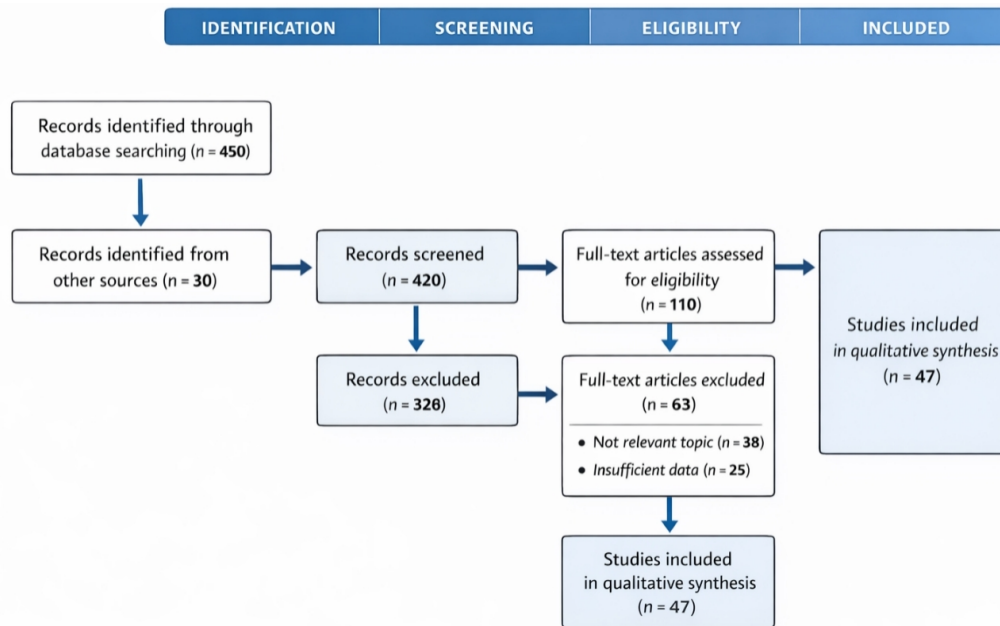


Figure 2. PRISMA 2020 Percussion Research Identification Flow Diagram.

This diagram represents the stepwise procedure in study selection, including the identification of studies, screening, assessment of eligibility, and eventual inclusion of 47 studies in the qualitative synthesis.

3.4. Data Extraction and Quality Assessment

Cahyaningtyas et al. [2] provided conventional measures for retrieving the performance data of deep learning models in segment-and-classify transcription jobs. Sections, downbeats, and microtiming are multiscale rhythmic frameworks. Ishizuka et al. [19] suggested the use of self-attention topics to retrieve and analyze the global form of transcribed drum sequences. Ishizuka et al. [17] explained that tatum-level information extraction can be trained on a regularized convolutional recurrent network. As Ni [4] explained, features based on STFT can be extracted to enhance the smart production of drum music scores. Jajoria and McDermott [40] evaluated the quality of symbolic drumbeats by providing features of text-conditioned latent diffusion models.

3.5. Quality Assessment

A quality assessment framework was employed to ensure reliability and validity of the included studies. The following criteria were used to assess each study:

1. Methodological transparency.

2. Description of the dataset and availability.
3. Model evaluation rigor.
4. Cultural contextualization.

A three-level scale was used to measure each criterion (low, medium, high). Studies that failed to comply with acceptable standards of quality were excluded from the ultimate synthesis. This would strengthen the validity and strength of the review findings.

3.6. Synthesis and Analytical Framework

Barnabò et al. [38] performed an example of this analysis by discussing the arrangement of drums created using CycleGAN frameworks and highlighting that this part of the synthesis process was essential. Oludare [41] explored the generative grammars of Yoruba timelines to validate the sociocultural significance of identified rhythmic patterns. Morais et al. [42] generalized the lessons learned as augmented meter-tracking models to explain the lack of certain signature times throughout the world in percussion traditions. McGraw and Benadon [30] combine both ethnographic and data-driven models with the view to model the situational experience of rhythm among Santiago de Cuba practitioners. Burloiu [31] introduced an interactive model that synthesizes the microtiming errors necessary to make expressive performances on a drum machine. A methodological framework that follows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guideline is outlined in **Table 2**, elaborating on the major phases, associated reference materials, and technical requirements necessary to conduct the synthesis.

Table 2. Methodological parameters of the PRISMA review.

PRISMA Stage	Author and ID Reference	Technical Requirement for Synthesis
Identification	The PRISMA method provides the initial search workflow.	Records were gathered from multiple academic databases.
Criteria	Grollmisch [6] defines the terms for low-resource audio.	Instrument specificity was the primary inclusion criterion.
Evaluation	Rohit and Rao [5] identified harmonic resonance markers.	Timbral complexity must be addressed using the chosen models.
Alignment	Ravikumar [14] demonstrated human-in-the-loop validation.	Machine representations should align with human musical judgments.

The PRISMA outline describes a four-step identification procedure that begins with deleting duplicate records of the retrieval output. Grollmisch [6] then suggested instrument-specific inclusion criteria for the screening stage, with an emphasis on Hindustani and Afro-Latin percussion instruments. Rohit and Rao [5] mentioned the technical requirements for categorizing strokes to be examined during the eligibility assessment. Batlle-Roca et al. [15] completed the process by ensuring that transparency and ethics were met with all integrated generative AI systems. It is an iterative selection procedure that ensures that the final selection represents the most plausible and culturally compatible advances in reducing research on computational percussion.

4. Results

A standardized PRISMA 2020 framework was used to locate and filter 480 primary records, which finally led to 47 articles that met the rigorous qualification for research on cross-cultural percussion. Civit et al. [39] present the coverage of the findings in clear domains namely, transcription, analysis, and generation as an example of a technological shift to culturally sensitive AI architectures. Batlle-Roca et al. [15] noted that model transparency is also evaluated alongside technical accuracy to address ethical issues that may be posed by music-generative systems.

4.1. Systematic Search and Selection Characteristics

Huang et al. [43] argue that due to the disjointed nature of world music databases, it means that interdisciplinary screening should be conducted in order to recognize works that do not follow standard western paradigms. Born [44] points out that the practical issues of music information retrieval diversification can be traced to the increased number of publications devoted to underrepresented instruments, such as the tabla and conga. According to Duan et al. [10], the special issue of TISMIR on cultural diversity has acted as a catalyst for quality research in this field since 2023. **Table 3** illustrates the spread of the included studies based on the focus of the instruments (tabla,

conga/Afro-Latin, and combined) and the activities of the research, highlighting the prevalence of transcription-related research.

Table 3. Present the distribution of the final included studies by instruments and tasks.

Research Task	Tabla Focus [5,22]	Conga/Afro-Latin Focus [12]	Combined/Other [2,38]
Transcription	12	8	5
Rhythmic Analysis	4	7	2
Generative Synthesis	3	4	3

4.2. Timbral Classification Accuracy for Tabla

Rohit and Rao [5] achieved a state-of-the-art F-score of 0.965 using a test set with a leave-one-table-out cross-validation scheme, demonstrating that convolutional recurrent neural network (CRNN) models can easily traverse harmonious overlaps in Hindustani concert audio. According to Shete and Deshmukh [22], rhythmic prediction with a multi-vent coded SVM classifier trained on spectral descriptors yields 93.33% accuracy. Vinod and Colaco [23] reported that the depiction of the architecture of deep learning 6-hidden layers is able to categorize the primary rhythm patterns with 97.50 accuracy. Cahyaningtyas et al. [2] compared multiple deep learning methods, achieving an average F1 macro of 85.32% on various datasets at 240 ms onset latency. The dataset of the STAR Drums presented by Weber et al. [3] can serve as a benchmark because it demonstrates that the model will perform better when trained on a wide spectrum of high-quality audio representations. According to Riley and Dixon [25], even with an increase in accuracy, percussion transcription remains behind the analysis of pitched instruments because such large and labelled datasets are rare. Both Melucci et al. [47] and Zehren et al. [27] have underlined the fact that artificial data and crowdsourced annotations are becoming compulsory elements in fulfilling the enormous data-hungry nature of modern automatic drum transcription (ADT) algorithms. Convolutional recurrent neural networks (CRNNs) are deep learning models that are better performers than the traditional ones, as shown in **Figure 3**.

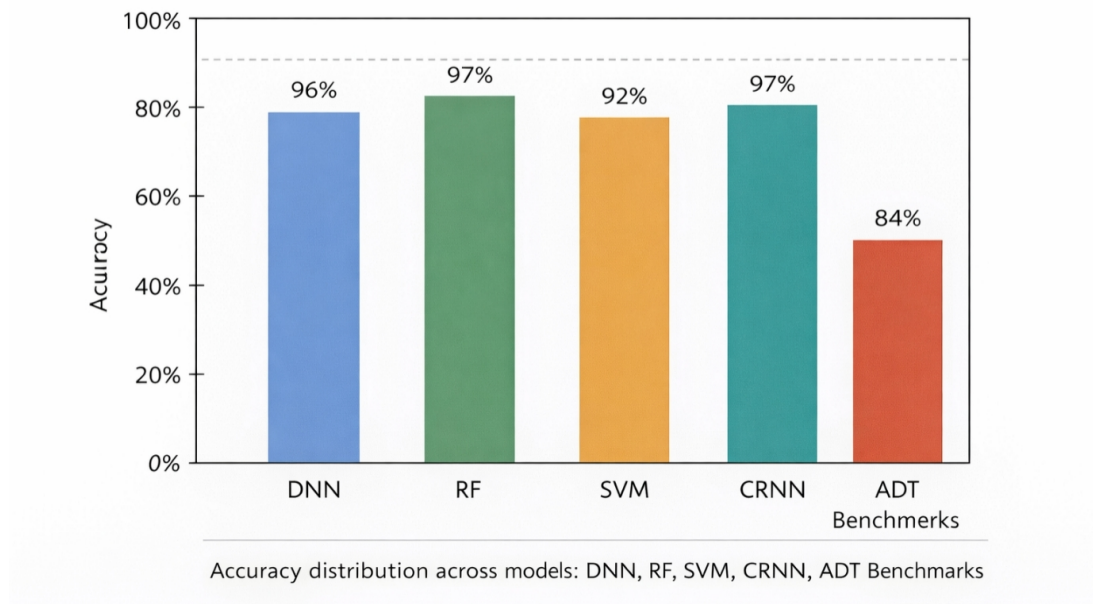


Figure 3. Tabla stroke classification against comparison.

The figure shows the classification performance depending on different modelling methods, including deep neural networks (DNNs), random forest (RF), support vector machines (SVMs), convolutional recurrent neural networks (CRNNs), and automatic drum transcription (ADT) benchmarks. Accuracy distribution across models: DNN, RF, SVM, CRNN, and ADT benchmarks.

4.3. Microtiming and Rhythmic Pattern Results

McGraw and Benadon [30] combined quantitative and ethnographic results and established the idea that rhythm is a way of filling spaces and not working on a strict grid of time. Danielsen et al. [34] discovered that acoustic features of time, such as duration and attack, significantly affect how listeners perceive such micro-rhythmic deviations. Burloiu [31] proposed an expressive drum machine that is interactive in learning the profile of micro-timing to play like a human. The SAMBASET dataset was provided by Maia et al. [33], enabling a large-scale analysis of the temporal structures of historical recordings of samba.

4.4. Low-Resource Meta-Learning Performance

The performance of meta-learning was tested by Grollmisch [6] in terms of the F1-score and a collar of 50 ms, and they showed that the performance of the PTM1 model exceeded other baselines by using the reference of the time context of all the frames in a stroke. Kodag and Arora [21] show that meta-learning in the form of optimization provides a quick adaptation of task and is significantly more resilient than the predetermined representation in cases of low-resource audio environments. As proven by Raimon et al. [1], meta-learning is similar to the process of human generalization, allowing models to be relevant to new percussion instruments with a minimum amount of data. Samuel et al. [13] successfully applied meta-learning extracts to the problem of music source separation and increased the applicability of the method to complex auditory decompositions. Shi et al. [36] note that the meta-based few-shot learning improves the detection of uncommon acoustic events thus boosting the overall performance of the system in general. Maia et al. [9] researched the possibility of reducing the labor costs of training these models to regional musical traditions by selective annotation plus onset detection of the human-in-the-loop.

4.5. Generative Synthesis and Audio Metrics

As shown by García et al. [12], when trained on augmented datasets, Monotonic Groove Transformers generate Afro-Cuban grooves that gradually become more faithful to real human performance. According to Barnabò et al. [38], the CycleDRUMS model, created by organizing drum tracks according to given bass lines, has an accuracy of 93% during user reviews. Jajoria and McDermott [40] evaluated the generation of symbolic drumbeats through latent diffusion models, and participants' feedback indicated a high rating of both novelty and contextual appropriateness. Oludare [41] used a generative approach in which she decoded the grammatical structures of Yoruba timelines and placed them into context as a form of social communication. Morais et al. [42] found that domain drift in generative systems can be alleviated by augmenting meter-tracking models to poorly represent time signatures. Ni [4] demonstrated that increased convolutional neural network designs, coupled with short-time Fourier transform features, are capable of producing drum music scores that possess both a high level of interpretive sophistication and style accuracy.

4.6. Human-in-the-Loop and Alignment Metrics

Ravikumar [14] provided the DAIRHuM platform, which ensures that human judgments with respect to music are directly correlated with the representations of artificial intelligence in Carnatic music. Porcaro et al. [45] indicated that AI-based systems used to create music may support cultural bias when they do not establish multicultural conceptions of diversity in their creation. Bai [46] utilized multimodal deep-learning models to combine machine pattern recognition with human perception of rhythm. As demonstrated by Muljono et al. [16], deep-learning methods for tempo recognition must be tailored to the particularities of audio signals to ensure precision in a wide variety of music types. Both Ishizuka et al. [19] and Ishizuka et al. [17] believed that tatum-level and symbol-level evaluations are essential to the overall evaluation of the transcription performance of a model in high-speed percussive applications.

5. Discussion

The discussion is organized based on the three dimensions of analysis introduced in the Introduction to synthesize the findings of the studies in a systematic manner and to be consistent with the review structure. The set of results synthesized between 2020 and 2026 can be characterized as a major shift in computational studies of percussion, moving from an inflexible frame-level classification to a culturally situational, meta-learning framework.

This change is a direct counter to the failure of Western-centric models when applied to other instruments that have unique sounds and rhythmic qualities in their musical profiles, making them unique musical entities.

5.1. Timbral Representation and Stroke Modeling

Rohit et al. [5] reveal that Convolutional Recurrent Neural Networks counter this acoustic masking of Hindustani concert audio and score 0.965 in the F-score. This finding is important because it proves that artificial intelligence can distinguish between complex harmonic and percussive overlaps. Shete and Deshmukh [22] further confirmed the same idea stating that spectral descriptors with multi-SVM classifiers achieve a 93.33% accuracy in predicting rhythmic cycles, and that spectral descriptors are more capable of capturing the syllabic quality of Tabla performance with machine-learning techniques. Vinod and Colaco [23] validate that deep neural networks can currently recognize the key beat cycles at a rate of 97.50% with proper training along with testable features reflecting resonant properties of the instrument.

5.2. Low-Resource Learning and Model Adaptability

Grollmisch [6] argues that meta-learning has been a turning point in the study of ethnomusicology as it enables the training of models on datasets, which are scarcely annotated. Their PTM1 model is designed to capitalize on the temporal context of the entire stroke segment, which, as Kodag and Arora [21] note, is more resistant to offline annotation errors or algorithms that rely only on onset frames, which are often vulnerable to human annotation errors. This methodology, which logs onto human generalization, is supported by Raimon et al. [1] to allow artificial intelligence to accommodate uncommon percussion instruments at a speed and accuracy that Samuel et al. [13] had already described as necessary for music source separation under low-resource constraints. This ability to adapt to diverse tasks is a proficient way of reducing the issue of the lack of data that has traditionally hindered the use of deep learning in different music traditions worldwide.

5.3. Cross-Cultural Transferability and Domain Adaptation

A key issue identified in this review is the low transferability of AI models to non-Western music when they are trained on Western percussion datasets. The majority of traditional models are formulated based on grid-based rhythmic organization and largely on inharmonic timbral features, which are insufficient for capturing the acoustic and rhythmic nature of instruments, such as the Hindustani tabla or Afro-Cuban conga.

This incompatibility causes a domain shift in which models degrade significantly when used in different cultural situations. Specifically, the hybrid harmonic-percussive quality of tabla strokes, as well as microtiming variants inherent in the structure of Afro-Latin beats, are challenging aspects that are not reflected in Western-centric training samples.

Transfer learning and meta-learning methods have become viable approaches to overcome such drawbacks. Meta-learning models exhibit better generalization to culturally diverse musical domains because of their ability to quickly adapt to new data distributions with only a small number of labeled samples. Likewise, those methods of data augmentation include pitch shifting, changing of time and creation of synthetic data, which, in turn, assist in decreasing data scarcity and enhancing robustness.

Nonetheless, the results indicate that technical adaptation is insufficient. To ensure true cross-cultural transferability, culturally informed representations, such as rhythm-specific systems (e.g., tala and clave), and knowledge of performance practice must be integrated. Without such integration, AI systems have a limited chance of producing technically correct but culturally inappropriate output.

Thus, future studies ought to give special emphasis to the hybrid methods involving the integration of machine learning and ethnomusicological understanding in order to guarantee the efficiency of the cross-cultural music information retrieval system along with the maintenance of authenticity and convention.

5.4. Rhythmic Structure and Microtiming Analysis

McGraw and Benadon [30] present quantitative evidence of such improvisational nuances not being well captured by conventional notation in the context of ethnographic studies, who discovered that rhythm is defined by practitioners as something that fills a space. Danielsen et al. [34] also argue that acoustic characteristics, namely

duration and attack, have a decisive effect on how micro-deviations are perceived by the listener and, therefore, can make the problem of AI models that begin with crude onset detection and do not consider culturally specific rhythmic models, such as the clave, more complex.

5.5. Generative Authenticity and Stylistic Data Augmentation

Garcia et al. [12] also show that data augmentation is an essential technique to train the generative model like ClaveNet to generate stylistically true Afro-Cuban grooves. This emphasis on authenticity is reflected in the work of Barnabò et al. [38], whose CycleDRUMS model has a 93% success rate in synthesizing drum sections that follow the harmonic and rhythmic prompts of a bass line. Jajoria and McDermott [40] have emphasized the potential of latent diffusion models to generate new percussive textures from textual inputs and have shown that these models can assist researchers in deciphering the latent grammars of rhythmic timelines represented by the Oludare [41]. According to Morais et al. [42], augmenting tracking models of underrepresented time signatures is required to ensure that generative systems without domain drift are used in situations outside of the West.

5.6. Ethical Transparency and Human-AI Alignment

Ravikumar [14] presents the DAIRHuM system designed to guarantee that the coverage of the concept of AI-generated representations is in line with personal musical opinions of the professionals working in the Carnatic tradition. Such an alignment is crucial because the absence of openness in generative AI, according to Batlle-Roca et al. [15], is one of the critical ethical concerns in modern musicology. As Porcaro et al. [45] and Huang et al. [43] warn, AI systems can reinforce Western-centric biases and reduce the cultural array of practices related to percussion practices of people of different cultures without human-in-the-loop validation and interdisciplinary critiques. Born [44] finds that the practical subtleties relating to the diversification of music information retrieval compel a basic reevaluation of the procedures through which knowledge is formed and expressed in the context of the given field.

5.7. Technical Bottlenecks and Future Research Directions

As Cahyaningtyas et al. [2] note, despite the fact that the recognition latency is already lowered to 240 ms, there are still significant technical challenges in the way of providing real-time transcription, in multi-phonetic acoustic conditions. Riley and Dixon [25] argue that this lack of variety and high quality data set is the main deterrent to the new developments in automatic drum transcription. According to Ishizuka et al. [19] and Ishizuka et al. [17], it is necessary to make developmental strides over and above the frame-, tatum-, and symbol-level measurements to have a complete picture of how artificial intelligence deals with high-speed and complex metrical structures of world percussion. Subsequent studies should thus focus on the creation of effective models that combine spectral accuracy with a culture-based understanding of rhythmic insight and be able to encompass the variety of world music.

6. Conclusion

Using the PRISMA 2020 guidelines, this systematic review summarizes the findings from 2020 to 2026 on the topic of the computational analysis of percussion. The record shows a strong shift from rigid Western metrical structures to models that are more culturally sensitive to the unique acoustic and rhythmic qualities of the tabla and conga. Recent improvements in deep learning design, especially convolutional recurrent neural networks, have enabled more accurate simulated Hindustani concert recording models than ever before. Simultaneously, the introduction of meta-learning models has partly circumvented the lack of data in ethnomusicology because it is now possible to quickly apply models to new instruments with only a small number of annotated samples required.

As we have seen in the analysis of rhythmic structure, systematic microtiming at the tatum level is a primary part of Afro-Latin musical identity and not an accidental variant. Suites of beats and microtiming profile monitoring mechanisms have become essential for capturing the inherent groove of such traditions. In the generative sphere, modern models achieve high accuracy by producing stylistically realistic accompaniments that follow systematic rhythmic grammars. However, salient technical barriers persist, such as in the area of real-time transcription in polyphonic situations and the long-term need for heterogeneous and high-integrity data.

Future research should focus on developing a platform that aligns AI-generated representations with professional human musical judgments, which would guarantee positive ethical advancements in music information search. Clear ethical concerns must remain at the forefront to protect the intellectual and cultural heritage represented. Finally, this study provides a strict basis for further studies in which cultural specificity and human-centered design of the further evolution of global percussion AI can be predicted.

Funding

This research received no external funding.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

The data presented in this study are available in this article. No new datasets were created or analyzed; the analysis was limited to the literature review.

Acknowledgments

The authors would like to acknowledge Srinakharinwirot University for its academic support.

Conflicts of Interest

The authors declare no conflict of interest.

AI Use Statement

During the preparation of this manuscript, the authors used Grammarly alone for language refinement. AI tools were not used for data analysis, interpretation, or the generation of scientific content. All outputs were critically reviewed and edited by the authors. The authors take full responsibility for the integrity and accuracy of this work.

References

1. Raimon, A.; Masti, S.; Sateesh, S.K.; et al. Meta-Learning in Audio and Speech Processing: An End to End Comprehensive Review. *arXiv preprint* **2024**, *arXiv.2408.10330*. [\[CrossRef\]](#)
2. Cahyaningtyas, Z.A.; Purwitasari, D.; Fatichah, C. Deep Learning Approaches for Automatic Drum Transcription. *EMITTER Int. J. Eng. Technol.* **2023**, *11*, 21–34. [\[CrossRef\]](#)
3. Weber, P.; Uhle, C.; Müller, M.; et al. STAR Drums: A Dataset for Automatic Drum Transcription. *Trans. Int. Soc. Music Inf. Retr.* **2025**, *8*, 248–264. [\[CrossRef\]](#)
4. Ni, Y. Intelligent generation method of drum music scores based on improved CNN and STFT. *Sci. Rep.* **2025**, *15*, 35322. [\[CrossRef\]](#)
5. Rohit, M.A.; Rao, P. Automatic Stroke Classification of Tabla Accompaniment in Hindustani Vocal Concert Audio. *arXiv preprint* **2022**, *arXiv.2104.09064*. [\[CrossRef\]](#)
6. Grollmisch, S. Semi-Supervised and Transfer Learning for Few-Shot Audio Classification. PhD Thesis, Technische Universität Ilmenau, Ilmenau, Germany, 2025.
7. Gowriprasad, R.; Kuriakose, J.; Srikrishnan, S.; et al. Unsupervised stroke clustering for developing common stroke set for percussion in Indian art music. *J. Acoust. Soc. Am.* **2025**, *158*, 3711–3723.
8. Damsma, A.; Cannon, J.; Fink, L.K.; et al. Computational modeling of rhythmic expectations: Perspectives, pitfalls, and prospects. *PLoS Comput. Biol.* **2025**, *21*, e1013798. [\[CrossRef\]](#)
9. Maia, L.S.; Rocamora, M.; Biscainho, L.W.P.; et al. Selective Annotation of Few Data for Beat Tracking of Latin American Music Using Rhythmic Features. *Trans. Int. Soc. Music Inf. Retr.* **2024**, *7*, 99–112. [\[CrossRef\]](#)

10. Duan, Z.; van Kranenburg, P.; Nam, J.; et al. Editorial for TISMIR Special Collection: Cultural Diversity in MIR Research. *Trans. Int. Soc. Music Inf. Retr.* **2023**, *6*, 203–205. [\[CrossRef\]](#)
11. Jiang, L.; Zhu, G.; Sun, J.; et al. Exploring the occupational biases and stereotypes of Chinese large language models. *Sci. Rep.* **2025**, *15*, 18777. [\[CrossRef\]](#)
12. García, D.F.; García, H.F.; Riondato, M. ClaveNet: Generating Afro-Cuban Drum Patterns through Data Augmentation. In Proceedings of the 19th International Audio Mostly Conference: Explorations in Sonic Cultures, Milan, Italy, 18–20 September 2024; pp. 355–361. [\[CrossRef\]](#)
13. Samuel, D.; Ganeshan, A.; Naradowsky, J. Meta-Learning Extractors for Music Source Separation. In Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 816–820. [\[CrossRef\]](#)
14. Ravikumar, P.T. DAIRHuM: A Platform for Directly Aligning AI Representations with Human Musical Judgments applied to Carnatic Music. *arXiv preprint* **2024**, *arXiv.2411.14907*. [\[CrossRef\]](#)
15. Batlle-Roca, R.; Gómez, E.; Liao, W.; et al. Transparency in Music-Generative AI: A Systematic Literature Review. *Research Square* **2023**. [\[CrossRef\]](#)
16. Muljono, M.; Andono, P.N.; Wulandari, S.A.; et al. Deep learning for audio signal-based tempo classification scenarios. *IAES Int. J. Artif. Intell.* **2024**, *13*, 1687–1687. [\[CrossRef\]](#)
17. Ishizuka, R.; Nishikimi, R.; Nakamura, E.; et al. Tatum-Level Drum Transcription Based on a Convolutional Recurrent Neural Network with Language Model-Based Regularized Training. *arXiv preprint* **2020**, *arXiv.2010.03749*. [\[CrossRef\]](#)
18. Smith, E.D. A New Foundation of Music: Sampling and Its Impact on the Creative Process. PhD Thesis, Princeton University, Princeton, NJ, USA, 2025.
19. Ishizuka, R.; Nishikimi, R.; Yoshii, K. Global Structure-Aware Drum Transcription Based on Self-Attention Mechanisms. *arXiv preprint* **2021**, *arXiv.2105.05791*. [\[CrossRef\]](#)
20. Zang, L.; Kim, H. Culture and the Social Clock: Cultural Differences in the Optimal Timing of Life. *Pers. Soc. Psychol. Bull.* **2025**, 01461672251362514. [\[CrossRef\]](#)
21. Kodag, R.B.; Arora, V. Meta-learning-based percussion transcription and Tāla identification from low-resource audio. *arXiv preprint* **2025**, *arXiv.2501.04742*. [\[CrossRef\]](#)
22. Shete, S.; Deshmukh, S. North Indian Classical Music Tabla Tala (Rhythm) Prediction System Using Machine Learning. In *Advances in Intelligent Systems and Computing*; Springer: Singapore, 2021; pp. 187–197.
23. Vinod, R.; Colaco, R. Review of AI in Indian Classical Music. *Int. J. Res. Appl. Sci. Eng. Technol.* **2025**, *13*, 1599–1608. [\[CrossRef\]](#)
24. Kaur, J.; Sohal, R.S.; Kaur, M.; et al. Acoustic Feature Synergy and Self-Supervised Learning for Robust Tabla Stroke Classification. *Research Square* **2026**. [\[CrossRef\]](#)
25. Riley, X.; Dixon, S. Enhanced Automatic Drum Transcription via Drum Stem Source Separation. *arXiv preprint* **2025**, *arXiv.2509.24853*. [\[CrossRef\]](#)
26. Moysis, L.; Iliadis, L.A.; Sotiroudis, S.P.; et al. Music Deep Learning: Deep Learning Methods for Music Signal Processing—A Review of the State-of-the-Art. *IEEE Access* **2023**, *11*, 17031–17052. [\[CrossRef\]](#)
27. Zehren, M.; Alunno, M.; Bientinesi, P. High-Quality and Reproducible Automatic Drum Transcription from Crowdsourced Data. *Signals* **2023**, *4*, 768–787. [\[CrossRef\]](#)
28. Hajek, J.C. Chocolate Surrealism: Music, Movement, Memory, and History in the Circum-Caribbean. *Ethnomusicology* **2024**, *68*, 164–166. [\[CrossRef\]](#)
29. Buisson, M.; McFee, B.; Essid, S.; et al. Self-Supervised Learning of Multi-Level Audio Representations for Music Segmentation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2024**, *32*, 2141–2152. [\[CrossRef\]](#)
30. McGraw, A.; Benadon, F. Representing and experiencing rhythm in drumming from Santiago de Cuba. In *Trends in World Music Analysis*; Routledge: London, UK, 2022; pp. 100–124.
31. Burloiu, G. Interactive Learning of Microtiming in an Expressive Drum Machine. In Proceedings of the 2020 Joint Conference on AI Music Creativity, Stockholm, Sweden, 21–24 October 2020.
32. Undy, C.; Dean, R.; Milne, A. An Algorithm for Human Embodiment and Computational Realization of Familiar and Novel African/Clave Rhythm. *Leonardo* **2025**, *59*, 301–309. [\[CrossRef\]](#)
33. Maia, L.S.; Fuentes, M.; Biscainho, L.W.P.; et al. SAMBASET: A Dataset of Historical Samba de Enredo Recordings For Computational Music Analysis. In Proceedings of the 20th International Society for Music Information Retrieval Conference, Delft, The Netherlands, 4–8 November 2019.
34. Danielsen, A.; Brøvig, R.; Böhler, K.K.; et al. There's More to Timing than Time: Investigating Musical Microrhythm across Disciplines and Cultures. *Music Percept.* **2024**, *41*, 176–198. [\[CrossRef\]](#)
35. Mnasri, Z.; Rovetta, S.; Masulli, F. Anomalous sound event detection: A survey of machine learning based

- methods and applications. *Multimed. Tools Appl.* **2022**, *81*, 5537–5586. [\[CrossRef\]](#)
36. Shi, B.; Sun, M.; Puvvada, K.C.; et al. Few-Shot Acoustic Event Detection Via Meta Learning. In Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 76–80. [\[CrossRef\]](#)
37. Pinto, A.S. Towards Human-in-the-Loop Onset Detection: A Transfer Learning Approach for Maracatu. *arXiv preprint* **2025**, *arXiv.2507.04858*. [\[CrossRef\]](#)
38. Barnabò, G.; Trappolini, G.; Lastilla, L.; et al. CycleDRUMS: Automatic drum arrangement for bass lines using CycleGAN. *Discover Artif. Intell.* **2023**, *3*, 4. [\[CrossRef\]](#)
39. Civit, M.; Civit-Masot, J.; Cuadrado, F.; et al. A systematic review of artificial intelligence-based music generation: Scope, applications, and future trends. *Expert Syst. Appl.* **2022**, *209*, 118190. [\[CrossRef\]](#)
40. Jajoria, P.; McDermott, J.B. Text Conditioned Symbolic Drumbeat Generation using Latent Diffusion Models. *arXiv preprint* **2024**, *arXiv.2408.02711*. [\[CrossRef\]](#)
41. Oludare, O. Towards a Generative Approach in Understanding the Kónkóló Timeline in Yoruba Music. *Balt. Int. Yearb. Cogn. Log. Commun.* **2023**, *15*, 2. [\[CrossRef\]](#)
42. Morais, G.; McFee, B.; Fuentes, M. Skip That Beat: Augmenting Meter Tracking Models for Underrepresented Time Signatures. *arXiv preprint* **2025**, *arXiv.2502.12972*. [\[CrossRef\]](#)
43. Huang, R.S.; Holzapfel, A.; Sturm, B.L.; et al. Beyond Diverse Datasets: Responsible MIR, Interdisciplinarity, and the Fractured Worlds of Music. *Trans. Int. Soc. Music Inf. Retr.* **2023**, *6*, 43–59. [\[CrossRef\]](#)
44. Born, G. Diversifying MIR: Knowledge and Real-World Challenges, and New Interdisciplinary Futures. *Trans. Int. Soc. Music Inf. Retr.* **2020**, *3*, 193–204. [\[CrossRef\]](#)
45. Porcaro, L.; Gómez, E.; Castillo, C. Perceptions of Diversity in Electronic Music: the Impact of Listener, Artist, and Track Characteristics. *Proc. ACM Hum.-Comput. Interact.* **2022**, *6*, 1–26. [\[CrossRef\]](#)
46. Bai, W. Pattern recognition study of musical rhythm perception based on multimodal deep learning. *Appl. Math. Nonlinear Sci.* **2024**, *9*, 1–14.
47. Melucci, P.; Merialdo, P.; Akama, T. Towards Realistic Synthetic Data for Automatic Drum Transcription. *arXiv preprint* **2026**, *arXiv.2601.09520*. [\[CrossRef\]](#)



Copyright © 2026 by the author(s). Published by UK Scientific Publishing Limited. This is an open access article under the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Publisher's Note: The views, opinions, and information presented in all publications are the sole responsibility of the respective authors and contributors, and do not necessarily reflect the views of UK Scientific Publishing Limited and/or its editors. UK Scientific Publishing Limited and/or its editors hereby disclaim any liability for any harm or damage to individuals or property arising from the implementation of ideas, methods, instructions, or products mentioned in the content.