VT INTERNATIONAL PRESS
VTIPJMRR | Journal of Multidisciplinary Research and Review

(Review Article)

Artificial Intelligence in Musical Instrument Learning: A Structured Review of Audio Signal Processing, Computer Vision, and Personalized Feedback

Vasudha Patil[1], **Mrunal Gholap**[2], **Anusha Bairagi**[3], **Tejaswini Vetal**[4]

[1]*Assistant Professor & HOD, Department of Artificial Intelligence and Data Science, Shri Chhatrapati Shivaji Maharaj College of Engineering, Ahilyanagar, Maharashtra, India*

[2,3,4] *BE Scholars, Department of Artificial Intelligence and Data Science, Shri Chhatrapati Shivaji Maharaj College of Engineering, Ahilyanagar, Maharashtra, India*

Corresponding Author: mrunal.aim@gmail.com

Received: 01. 09.2026	Accepted: 20.09.2026	Published: 28.09.2026

ABSTRACT

Artificial intelligence (AI) is increasingly used to support learners who practise a musical instrument without a teacher present. This structured narrative review examines how audio signal processing, computer vision, and their combination are used to assess instrumental performance, and how the resulting assessments are turned into personalized feedback. Literature was identified through Google Scholar and publisher databases, the ISMIR and NIME proceedings archives, and backward and forward citation chasing, with a final search in October 2026; 45 sources were retained and synthesized thematically against six research questions. Audio methods for monophonic pitch and onset estimation are mature and well benchmarked, whereas polyphonic transcription and guitar-specific string–fret estimation remain difficult. Camera-based hand-landmark tracking is now accessible on consumer devices, but landmarks do not by themselves establish correct fingering; instrument-specific studies remain small and are rarely tested outside the laboratory. Audio–visual fusion is promising because each modality resolves ambiguities in the other, yet few studies compare fused and single-modality systems on identical test material. Evidence on learning outcomes is thinner still: most systems report technical accuracy, and only a handful report controlled learner studies. The review closes with an evaluation framework that separates technical, pedagogical, and usability claims, and applies it to VIBEX, a guitar-focused learning platform under development by the authors.

Keywords: *artificial intelligence; music education; instrument learning; computer vision; audio signal processing; multimodal performance assessment; personalized feedback*

I. INTRODUCTION

Learning a musical instrument depends on sustained, goal-directed practice in which the learner repeatedly attempts a task, notices errors, and adjusts (Ericsson et al., 1993). In lessons, a teacher supplies much of this error detection: they listen to the sound, watch the hands and posture, diagnose the cause of a problem, and choose the next exercise. Between lessons, however, most practice happens alone, and the learner must judge their own performance. Percival et al. (2007) argued that this unsupervised daily practice, rather than the lesson itself, is where computer assistance is most needed, because what the learner lacks is objective and timely feedback. Research on feedback in education supports this view: feedback is among the strongest influences on learning, provided it tells learners where they are going, how they are doing, and what to do next (Hattie & Timperley, 2007).

AI-assisted systems can supply parts of this feedback through two complementary channels. Audio analysis can estimate pitch, note onsets, and rhythm, while computer vision can estimate hand landmarks, finger placement, and posture. An instructional layer can then compare these estimates with the target exercise and present corrections. Each step, however, introduces its own uncertainty. A detected hand landmark is not the same as a correctly fretted note, and an accurate pitch estimate says nothing about tone, phrasing, or whether the learner used an efficient fingering. Lerch et al. (2020) note that performance analysis spans acoustics, music psychology, and pedagogy, and that technical measurements must be interpreted against musically meaningful criteria before they can guide a learner.

Earlier reviews have mapped parts of this territory. Zhang et al. (2024) reviewed AI in music teaching and learning and identified personalization, timely feedback, interactivity, and organized digital resources as recurring benefits.

 © 2026 VT International Press. All rights reserved. E-ISSN: 3139-7026

Mohammadian et al. (2026) reviewed technology-enhanced music-learning models and found that performance proficiency was the most frequently targeted outcome, while pedagogical challenges were reported more often than technological ones. On the technical side, Tavares et al. (2013) and Benetos et al. (2019) surveyed automatic music transcription, Duan et al. (2019) reviewed audio-visual analysis of music performances, and Eremenko et al. (2020) assessed the readiness of performance-assessment technology for instrument teaching. These reviews address either the educational or the technical side; none brings audio analysis, computer vision, multimodal fusion, and feedback design together around the practical question of how a practice-support system should be built and evaluated.

This review addresses that gap through six research questions (RQs): (RQ1) Which audio methods are used to assess pitch, notes, rhythm, and timing in instrument learning? (RQ2) Which visual methods are used to assess hand position, fingering, posture, or technique? (RQ3) How are audio and visual evidence combined, and with what benefit? (RQ4) What feedback and personalization strategies are used, and how are they evaluated? (RQ5) What limitations are reported concerning datasets, robustness, latency, accessibility, and usability? (RQ6) How can these findings inform a testable evaluation plan for an AI-assisted guitar-learning platform?

The contribution of the paper is threefold: a thematic synthesis of the technical and pedagogical literature (Section III); a characteristics table of representative primary studies (Table 3); and an evaluation framework that separates technical, pedagogical, and usability claims (Section IV, Table 4), illustrated with VIBEX, a guitar-learning platform under development by the authors. Section II describes the review method, Section III reports the findings by research question, Section IV discusses gaps and implications, and Section V concludes.

II. REVIEW METHOD

A. Review Design

The study is a structured narrative review in the sense of Grant and Booth (2009): it uses an explicit search strategy, stated eligibility criteria, and a defined synthesis procedure, but it does not claim the exhaustive, dual-screened coverage of a systematic review. This design was chosen because the relevant evidence is spread across signal processing, computer vision, human–computer interaction, and education, and is published in venues (conference proceedings, datasets, technical reports) that bibliographic databases index unevenly. Reporting follows the PRISMA 2020 items that apply to this design, namely information sources, search strategy, eligibility criteria, data items, and synthesis method (Page et al., 2021). Because screening was not performed in duplicate with recorded counts at each stage, no PRISMA flow diagram is reported, and no pooled quantitative estimates are calculated.

B. Information Sources and Search Strategy

Searches were carried out through Google Scholar and general web search, and candidate records were checked against publisher and archive pages in IEEE Xplore, the ACM Digital Library, SciTePress, MDPI, MIT Press, and arXiv, together with the proceedings archives of the International Society for Music Information Retrieval (ISMIR) and the International Conference on New Interfaces for Musical Expression (NIME). The final search was run in October 2026. Search strings combined one term from each of three concept blocks (Table 1) and were adapted to each source. Reference lists of included reviews were screened (backward citation chasing), and citing articles of key technical papers were inspected (forward citation chasing).

Table 1. Search concept blocks. One term from each block was combined with AND; terms within a block were combined with OR.

Concept block	Example terms
Learning context	music education; instrument learning; music practice; music tutoring; performance assessment; guitar OR piano OR violin OR flute
Technique	artificial intelligence; machine learning; deep learning; pitch estimation; onset detection; automatic music transcription; hand tracking; pose estimation; fingering recognition; audio-visual; multimodal
Outcome / function	feedback; personalized; adaptive; intelligent tutoring; assessment; evaluation; usability; latency

C. Eligibility Criteria

Eligibility criteria were fixed before screening and are listed in Table 2. Foundational technical papers (for example, on pitch estimation or hand tracking) were included when their method is used directly in instrument-

 © 2026 VT International Press. All rights reserved. E-ISSN: 3139-7026

learning or performance-assessment systems, so that the review can explain what those systems inherit from their building blocks.

Table 2. Eligibility criteria.

Inclusion	Exclusion
Peer-reviewed journal articles, conference papers, and dataset papers on instrument learning, performance assessment, or AI-supported music education.	Duplicate records and later versions superseded by a peer-reviewed publication.
Foundational audio or vision methods that are directly used in instrument-learning or assessment systems.	Music generation or recommendation papers with no link to learning, assessment, or feedback.
Educational and psychological work on feedback, practice, tutoring, or usability that informs the design of feedback.	Promotional material or commercial product descriptions without methodological detail.
Preprints only where they describe a widely used method or dataset and no peer-reviewed version yet exists.	Records whose full text could not be accessed.
Written in English; no lower date limit for foundational work, emphasis on 2014–2026 for applied systems.	Studies of non-instrumental music activities only (e.g., listening, composition) without performance assessment.

D. Data Extraction

For each included primary study, the following items were recorded: instrument; input modality (audio, video, depth, or wearable sensor); assessment task (e.g., pitch, onset, chord, fingering, bowing); data source and size; participants, where applicable; evaluation metric; reported outcome; and stated limitations. For reviews, the scope, search period, and main conclusions were recorded. Extraction was carried out by the authors and checked against the original publication.

E. Quality Appraisal

Formal risk-of-bias tools are designed for intervention studies and fit poorly with signal-processing and computer-vision papers. Each primary study was therefore appraised against five reporting criteria derived from Lerch et al. (2020) and Eremenko et al. (2020): (i) the assessment task is defined precisely; (ii) the dataset or participant sample is described; (iii) the reference annotation or ground truth is stated; (iv) the metric is defined and appropriate to the task; and (v) test conditions, including hardware and environment, are reported. Learner studies were additionally checked for a comparison condition and for ethical approval. These criteria inform the strength of the statements in Section III rather than serving as an exclusion threshold.

F. Synthesis

Findings were synthesized thematically (Braun & Clarke, 2006). Extracted items were coded deductively against RQ1–RQ5, and codes that did not fit an existing question (for example, dataset availability) were grouped into the limitations theme. Quantitative results are reported only as given by each study; because tasks, instruments, and metrics differ, no cross-study ranking or meta-analysis is attempted. Of the 45 sources retained, 13 are primary studies or datasets on instrument-learning systems (Table 3), 13 are foundational audio or vision methods and tools, 7 are reviews or overviews, and 12 concern feedback, practice, usability, and review methodology.

III. FINDINGS

A. Audio Signal Processing for Performance Assessment (RQ1)

Audio-based assessment rests on several distinct tasks that differ in difficulty and should be evaluated separately. Fundamental-frequency estimation for single notes is the most mature. The YIN algorithm (de Cheveigné & Kawahara, 2002) refined autocorrelation with a cumulative mean-normalized difference function and remains a common baseline; pYIN added probabilistic thresholds and hidden-Markov-model smoothing to reduce octave and voicing errors (Mauch & Dixon, 2014). Data-driven estimators followed: CREPE applies a convolutional network directly to the waveform and reported higher accuracy and noise robustness than pYIN (Kim et al., 2018), while SPICE learns pitch without labelled data through self-supervision (Gfeller et al., 2020). Open libraries such as librosa make these methods accessible to small development teams (McFee et al., 2015).

 © 2026 VT International Press. All rights reserved. E-ISSN: 3139-7026

Timing assessment depends on onset detection, for which Bello et al. (2005) set out the standard pipeline of an onset-detection function followed by peak picking. Full automatic music transcription (AMT), which recovers every note in polyphonic audio, is considerably harder. Tavares et al. (2013) and Benetos et al. (2019) both describe polyphonic AMT as an open problem, particularly when notes overlap in frequency or several instruments sound together. Progress has been fastest for piano, where joint prediction of onsets and frames substantially improved note-level accuracy (Hawthorne et al., 2018).

Guitar poses a specific problem: the same pitch can be played at several string–fret positions, so a correct pitch does not reveal which fingering was used. GuitarSet provides about three hours of acoustic guitar recordings by six players, annotated with string and fret positions using a hexaphonic pickup (Xi et al., 2018). It has enabled tablature estimation from audio with convolutional networks (Wiggins & Kim, 2019) and, more recently, higher-resolution guitar transcription through domain adaptation (Riley et al., 2024). In an educational setting, Eremenko et al. (2020) built a proof-of-concept assessment system for guitar exercises and concluded that the accuracy of sound-based assessment in practice apps still fell short of what professional teaching requires. Pati et al. (2018) showed that deep networks can predict expert ratings of student performances from audio, which moves assessment beyond note correctness towards musical quality, although such ratings depend on the rubric and raters used.

Across these studies, results are reported with different metrics: pitch error in cents, note-level precision, recall, and F-measure, onset deviation in milliseconds, and correlation with expert ratings. Standard implementations such as mir_eval (Raffel et al., 2014) make results comparable only when the same metric, tolerance window, and test set are used. For a learning tool, the relevant evaluation must also state whether input is monophonic or polyphonic, the recording device and environment, and the level of background noise.

B. Computer Vision for Hand Position and Technique (RQ2)

General-purpose pose estimation has made camera-based hand tracking widely available. OpenPose estimates body and hand keypoints for multiple people in real time (Cao et al., 2021), and MediaPipe Hands predicts 21 three-dimensional hand landmarks from a single RGB camera on mobile hardware (Zhang et al., 2020), within the broader MediaPipe framework for perception pipelines (Lugaresi et al., 2019). These tools are a foundation rather than an assessment: they locate fingers in the image but do not know where the strings, frets, or keys are.

Instrument-specific work adds that knowledge. For guitar, Burns and Wanderley (2006) demonstrated retrieval of fretting-hand fingering from video, and Cicconet et al. (2010) identified chords from fingertip positions on the fretboard with accuracy comparable to an audio method while being less sensitive to strumming noise. Marullo et al. (2026) combined MediaPipe landmarks with depth images to recognize four guitar chords plus an unknown class, reporting a balanced accuracy above 85% on unseen data from 18 participants recorded under the same conditions. For piano, Akbari and Cheng (2015) transcribed performances from an overhead camera alone, reporting an F-score of 0.974 for pressed-key detection with low latency, and Koepke et al. (2020) trained an end-to-end network to predict note onsets from video. Large datasets are now appearing: PianoMotion10M contains about 116 hours of overhead piano video with roughly ten million annotated hand poses (Gan et al., 2025), and PianoVAM pairs video, audio, MIDI, hand landmarks, and fingering labels from amateur pianists (Kim et al., 2025).

Three limitations recur. First, studies are small and recorded under controlled conditions; performance under different lighting, camera angles, hand sizes, and instruments is rarely tested. Second, the fretting hand frequently occludes itself, and the neck of the guitar may leave the frame, so tracking failures must be detected and handled rather than silently passed to the assessment stage. Third, most studies stop at recognition accuracy; few test whether the visual estimate is accurate enough to support a correct teaching decision, such as telling a learner which finger to move.

C. Audio–Visual Integration (RQ3)

Audio and video carry different evidence about a performance. Audio reveals the acoustic result but not its physical cause; video reveals the movement but not the sound produced. Duan et al. (2019) argue that this complementarity is the main motivation for multimodal analysis and that fusing low-level representations early is generally preferable to merging the outputs of two independent systems, although early fusion is harder because not every visible movement is linked to the sound. Datasets such as URMP, which provides separately recorded audio and video tracks of classical ensemble pieces, were created specifically to support this kind of research (Li et al., 2019).

Direct evidence that fusion improves assessment is still limited. Koepke et al. (2020) reported that combining their visual transcription with audio improved results relative to either modality alone, and Cicconet et al. (2010)

 © 2026 VT International Press. All rights reserved. E-ISSN: 3139-7026

showed that video can disambiguate chords where audio is degraded by strumming noise. For guitar learning, the potential benefit is clear in principle: audio can confirm that the intended pitch sounded, and vision can identify which string and fret produced it, resolving the string–fret ambiguity noted in Section III-A. However, no study identified in this review compared audio-only, vision-only, and fused guitar assessment on the same set of learner recordings, and the added computation, camera calibration, synchronization, and privacy costs of video are rarely weighed against the gain.

D. Feedback and Personalization (RQ4)

Detecting an error is only the first half of tutoring. Hattie and Timperley (2007) distinguish feedback about the task, the process, self-regulation, and the self, and find the first three most useful. Shute (2008) adds that formative feedback is most effective when it is specific and focused, and that elaborated feedback explaining how to correct an error usually outperforms simple right-or-wrong messages. For motor skills, Wulf and Shea (2002) caution that very frequent feedback can improve performance during practice while weakening retention, which argues for fading feedback as competence grows. More broadly, VanLehn (2011) found that step-based intelligent tutoring systems approach the effectiveness of human tutoring, which suggests that fine-grained, step-level feedback is a sensible design target.

Music-specific systems apply these ideas with varying rigour. Early projects such as i-Maestro combined gesture capture, augmented instruments, and multimedia to support practice and theory learning (Ng & Nesi, 2008), and Dittmar et al. (2012) described how music information retrieval techniques could underpin practice software. Controlled learner studies are rarer. Xia et al. (2020) evaluated an interactive, colour-coded flute tutor and reported that learners using the interactive version learned about 31% faster, with the largest benefit in early training. Provenzale et al. (2024) gave 24 beginner violinists either real-time visual feedback on bow angle from wearable inertial sensors or no feedback, and found improved bow control in the feedback group. Personalization in the reviewed literature mostly means adapting exercise difficulty or sequence to observed performance (Zhang et al., 2024); adaptive policies are seldom evaluated against a fixed curriculum.

E. Robustness, Latency, Accessibility, and Evaluation (RQ5)

The limitations reported across the literature cluster into five themes. **Datasets.** Annotated data with string–fret, fingering, or expert-rating labels is scarce, small, and often recorded by skilled players rather than learners (Xi et al., 2018; Lerch et al., 2020). **Robustness.** Most systems are tested in quiet rooms with fixed cameras, whereas home practice involves background noise, variable lighting, and inexpensive microphones (Eremenko et al., 2020). **Latency.** Interactive feedback requires low end-to-end delay; Akbari and Cheng (2015) report latency for a vision-only piano system, but few studies report the full input-to-feedback delay together with hardware specifications. **Usability and accessibility.** Usability is commonly measured with the System Usability Scale (Brooke, 1996), for which normative benchmarks exist (Bangor et al., 2008); accessibility for learners with motor or visual impairments is almost never examined. **Learning outcomes.** Mohammadian et al. (2026) report that pedagogical challenges outweigh technical ones in the literature they reviewed, and the present review found few studies with a comparison group, a retention test, or reported ethical approval.

Table 3. Characteristics of representative primary studies and datasets on instrument-learning and performance-assessment systems.

Study	Instrument	Modality	Task and data	Reported outcome
Burns & Wanderley (2006)	Guitar	Video	Retrieval of fretting-hand fingering	Feasibility of camera-based fingering retrieval demonstrated
Cicconet et al. (2010)	Guitar	Video	Chord identification from fingertip positions	Comparable to an audio method; more robust to strumming noise
Akbari & Cheng (2015)	Piano	Video	Pressed-key detection and transcription	F-score 0.974; low latency
Xi et al. (2018)	Guitar	Audio (hexaphonic)	Dataset: 180 excerpts, 6 players, ~3 h, string–fret labels	Benchmark for guitar transcription
Pati et al. (2018)	Wind and percussion (student)	Audio	Prediction of expert ratings of student performances	Deep networks predicted ratings from audio
Wiggins & Kim (2019)	Guitar	Audio	Tablature estimation with a CNN, trained on GuitarSet	String–fret estimation from audio shown feasible
Eremenko et al. (2020)	Guitar	Audio	Assessment of guitar exercises (proof of concept)	Accuracy not yet adequate for professional teaching

 © 2026 VT International Press. All rights reserved. E-ISSN: 3139-7026

Koepke et al. (2020)	Piano	Video (+ audio)	End-to-end visual onset transcription	Combining with audio improved results
Xia et al. (2020)	Flute	Visual notation, interaction	Learner study of an interactive tutor	About 31% faster learning with interaction
Provenzale et al. (2024)	Violin	Wearable inertial sensors	Real-time bowing feedback; 24 beginners (12 vs 12)	Improved bow control with feedback
Gan et al. (2025)	Piano	Video	Dataset: ~116 h, ~10 million hand poses	Benchmark for hand-motion modelling
Kim et al. (2025)	Piano	Video, audio, MIDI	Dataset with hand landmarks and fingering labels	Multimodal benchmark from amateur players
Marullo et al. (2026)	Guitar	RGB-D + MediaPipe	4 chords + unknown; 18 participants	Balanced accuracy above 85% on unseen data

IV. DISCUSSION

A. Synthesis Across Research Questions

Taken together, the literature describes instrument-learning support as a chain of dependent stages: capturing the signal, estimating musical or movement features, interpreting them against the exercise, and communicating a useful correction (Figure 1). Reliability decreases along this chain. Monophonic pitch and onset estimation (RQ1) and generic hand-landmark detection (RQ2) are well developed and well benchmarked; instrument-specific interpretation, such as identifying the string and fret or judging technique, is supported by far smaller and less varied evidence; and the final pedagogical stage (RQ4) is the least studied of all. Success at an upstream stage therefore cannot be taken as evidence of success downstream. A system that detects pitch accurately in clean, single-note recordings may still mislead a learner who plays chords in a noisy room, and reliable landmarks may still be insufficient to decide whether the intended string was pressed.

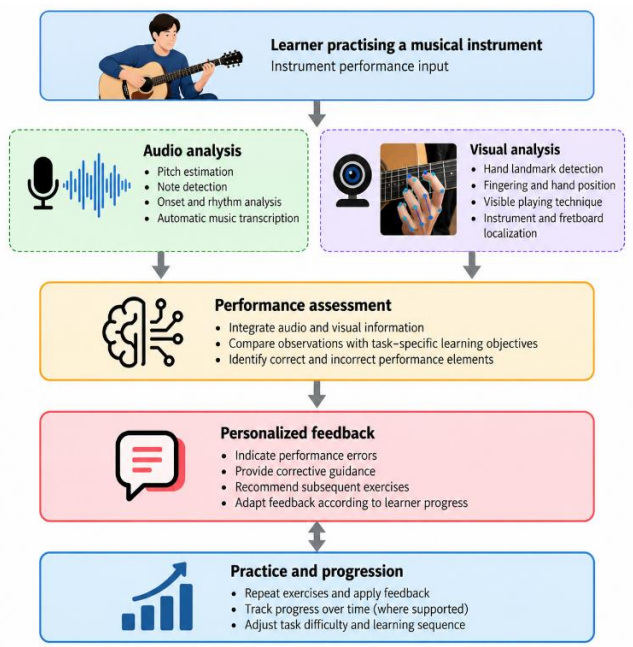


Figure 1. Conceptual framework of AI-assisted instrument learning derived from the review: audio and visual analysis feed a performance-assessment stage, which drives personalized feedback and practice progression.

Multimodal fusion (RQ3) is the most promising route to resolving instrument-specific ambiguities, but the claim that it improves assessment rests on a small number of studies, mostly on piano. The main gaps identified are: (i) the absence of learner-recorded, string–fret-annotated audio-visual guitar data; (ii) the lack of head-to-head comparisons of audio-only, vision-only, and fused assessment on identical material; (iii) inconsistent reporting of latency and test conditions; and (iv) the scarcity of controlled learner studies with retention measures and reported ethical approval (RQ5).

 © 2026 VT International Press. All rights reserved. E-ISSN: 3139-7026

B. Implications for the Design and Evaluation of VIBEX (RQ6)

VIBEX is a web-based guitar-learning platform under development by the authors, with Python services for audio analysis and camera-based hand tracking, structured lessons, and learner progression. Its implementation and evaluation will be reported in a separate empirical paper; here, the review findings are used to define how that evaluation should be designed. Three principles follow from Section III. First, the platform should begin with a narrow, testable use case, such as single-note and open-chord exercises, before claiming broader assessment. Second, each component should be validated on its own before any combined claim is made, and fusion should be retained only if it measurably improves assessment over the better single modality on the same test items. Third, technical accuracy, usability, and learning outcomes are different claims that require different evidence and should not be summarized as a single accuracy figure. Table 4 sets out the resulting evaluation framework.

Table 4. Evaluation framework derived from the review and proposed for VIBEX.

Component	Claim to be tested	Reference standard and metric	Minimum reporting
Pitch and note detection	The played note is identified correctly	Annotated notes; pitch error (cents), note-level precision, recall, F-measure (Raffel et al., 2014)	Microphone, guitar type, noise level, number of notes and players
Onset and timing	Note timing relative to the target is measured correctly	Annotated onsets; onset deviation (ms), F-measure with stated tolerance (Bello et al., 2005)	Tempo range, tolerance window
Hand tracking	Fretting-hand landmarks are located reliably	Manually annotated keypoints; detection rate, error normalized by hand size	Camera position, lighting, occlusion cases
String–fret and chord recognition	The fingering used is identified	Expert labels; per-class precision and recall, confusion matrix	Players not seen in training; failure cases
Audio–visual fusion	Fusion improves on the better single modality	Same test items for all three conditions; paired comparison with confidence intervals	Synchronization method, added computation
End-to-end latency	Feedback is delivered quickly enough for practice	Hardware time-stamps; median and 95th-percentile delay	Device specification, repeated trials
Usability	Learners find the system usable	System Usability Scale (Brooke, 1996), interpreted with Bangor et al. (2008)	Sample size, learner experience levels
Learning outcomes	Practice with feedback improves performance	Pre-test, post-test, and delayed retention test against a comparison group (Wulf & Shea, 2002)	Ethics approval, informed consent, recruitment, test instrument

C. Limitations of This Review

As a structured narrative review, this study does not guarantee exhaustive coverage. Screening was not performed in duplicate, record counts at each stage were not logged, and relevant work in languages other than English or in venues not indexed by the sources used may have been missed. The quality appraisal used reporting criteria rather than a validated risk-of-bias tool. The evidence base is also uneven: piano and guitar dominate, and several recent datasets are available only as preprints. Finally, because the authors are developing VIBEX, the implications in Section IV-B were drafted with that platform in mind; they are offered as an evaluation plan, not as evidence of its performance.

V. CONCLUSION

AI can already provide reliable building blocks for instrument-learning support: monophonic pitch and onset estimation, and camera-based hand tracking on everyday devices. The weaker links are instrument-specific interpretation, the combination of audio and visual evidence, and above all the translation of measurements into feedback that demonstrably helps learners improve. Future work should prioritize learner-recorded, richly annotated audio-visual datasets; head-to-head comparisons of single-modality and fused assessment; consistent reporting of latency and test conditions; and controlled learner studies with retention measures and ethical

 © 2026 VT International Press. All rights reserved. E-ISSN: 3139-7026

oversight. The evaluation framework in Table 4 offers a practical template for such work and will guide the forthcoming evaluation of the VIBEX platform.

ACKNOWLEDGMENT

The authors thank their project guide, Dr. Vasudha Patil and the faculty of the Department of Artificial Intelligence and Data Science, Shri Chhatrapati Shivaji Maharaj College of Engineering, Ahilyanagar, for their guidance and feedback.

REFERENCES

Akbari, M., & Cheng, H. (2015). Real-time piano music transcription based on computer vision. *IEEE Transactions on Multimedia, 17*(12), 2113–2121. https://doi.org/10.1109/TMM.2015.2473702

Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An empirical evaluation of the System Usability Scale. *International Journal of Human–Computer Interaction, 24*(6), 574–594. https://doi.org/10.1080/10447310802205776

Bello, J. P., Daudet, L., Abdallah, S., Duxbury, C., Davies, M., & Sandler, M. B. (2005). A tutorial on onset detection in music signals. *IEEE Transactions on Speech and Audio Processing, 13*(5), 1035–1047. https://doi.org/10.1109/TSA.2005.851998

Benetos, E., Dixon, S., Duan, Z., & Ewert, S. (2019). Automatic music transcription: An overview. *IEEE Signal Processing Magazine, 36*(1), 20–30. https://doi.org/10.1109/MSP.2018.2869928

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Brooke, J. (1996). SUS: A "quick and dirty" usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester, & I. L. McClelland (Eds.), *Usability evaluation in industry* (pp. 189–194). Taylor & Francis.

Burns, A.-M., & Wanderley, M. M. (2006). Visual methods for the retrieval of guitarist fingering. In *Proceedings of the International Conference on New Interfaces for Musical Expression* (pp. 196–199). https://doi.org/10.5281/zenodo.1176850

Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., & Sheikh, Y. (2021). OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 43*(1), 172–186. https://doi.org/10.1109/TPAMI.2019.2929257

Cicconet, M., Carvalho, P., Velho, L., & Gattass, M. (2010). Visual pitch class profile: A video-based method for real-time guitar chord identification. In *Proceedings of the International Conference on Computer Vision Theory and Applications (VISAPP)* (Vol. 2, pp. 487–490). SciTePress. https://doi.org/10.5220/0002824904870490

de Cheveigné, A., & Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America, 111*(4), 1917–1930. https://doi.org/10.1121/1.1458024

Dittmar, C., Cano, E., Abeßer, J., & Grollmisch, S. (2012). Music information retrieval meets music education. In M. Müller, M. Goto, & M. Schedl (Eds.), *Multimodal music processing* (Dagstuhl Follow-Ups, Vol. 3, pp. 95–120). Schloss Dagstuhl–Leibniz-Zentrum für Informatik. https://doi.org/10.4230/DFU.Vol3.11041.95

Duan, Z., Essid, S., Liem, C. C. S., Richard, G., & Sharma, G. (2019). Audiovisual analysis of music performances: Overview of an emerging field. *IEEE Signal Processing Magazine, 36*(1), 63–73. https://doi.org/10.1109/MSP.2018.2875511

Eremenko, V., Morsi, A., Narang, J., & Serra, X. (2020). Performance assessment technologies for the support of musical instrument learning. In *Proceedings of the 12th International Conference on Computer Supported Education (CSEDU)* (Vol. 1, pp. 629–640). SciTePress. https://doi.org/10.5220/0009817006290640

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review, 100*(3), 363–406. https://doi.org/10.1037/0033-295X.100.3.363

Gan, Q., Wang, S., Wu, S., & Zhu, J. (2025). PianoMotion10M: Dataset and benchmark for hand motion generation in piano performance. In *Proceedings of the International Conference on Learning Representations (ICLR)*. https://doi.org/10.48550/arXiv.2406.09326

Gfeller, B., Frank, C., Roblek, D., Sharifi, M., Tagliasacchi, M., & Velimirović, M. (2020). SPICE: Self-supervised pitch estimation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28*, 1118–1128.

Grant, M. J., & Booth, A. (2009). A typology of reviews: An analysis of 14 review types and associated methodologies. *Health Information & Libraries Journal, 26*(2), 91–108. https://doi.org/10.1111/j.1471-1842.2009.00848.x

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. https://doi.org/10.3102/003465430298487

Hawthorne, C., Elsen, E., Song, J., Roberts, A., Simon, I., Raffel, C., Engel, J., Oore, S., & Eck, D. (2018). Onsets and frames: Dual-objective piano transcription. In *Proceedings of the 19th International Society for Music Information Retrieval Conference* (pp. 50–57). https://doi.org/10.5281/zenodo.1492341

 © 2026 VT International Press. All rights reserved. E-ISSN: 3139-7026

Kim, J. W., Salamon, J., Li, P., & Bello, J. P. (2018). CREPE: A convolutional representation for pitch estimation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 161–165). https://doi.org/10.1109/ICASSP.2018.8461329

Kim, Y., Park, J., Bae, J., Kim, K., Kwon, T., Lerch, A., & Nam, J. (2025). *PianoVAM: A multimodal piano performance dataset* (arXiv:2509.08800). arXiv. https://doi.org/10.48550/arXiv.2509.08800

Koepke, A. S., Wiles, O., Moses, Y., & Zisserman, A. (2020). Sight to sound: An end-to-end approach for visual piano transcription. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Lerch, A., Arthur, C., Pati, A., & Gururani, S. (2020). An interdisciplinary review of music performance analysis. *Transactions of the International Society for Music Information Retrieval*, *3*(1), 221–245. https://doi.org/10.5334/tismir.53

Li, B., Liu, X., Dinesh, K., Duan, Z., & Sharma, G. (2019). Creating a multitrack classical music performance dataset for multimodal music analysis: Challenges, insights, and applications. *IEEE Transactions on Multimedia, 21*(2), 522–535. https://doi.org/10.1109/TMM.2018.2856090

Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.-L., Yong, M. G., Lee, J., Chang, W.-T., Hua, W., Georg, M., & Grundmann, M. (2019). *MediaPipe: A framework for building perception pipelines* (arXiv:1906.08172). arXiv. https://doi.org/10.48550/arXiv.1906.08172

Marullo, G., Amati, D., Barbarino, S., D'Onofrio, M., Sechi, T., Innocente, C., Vezzetti, E., & Ulrich, L. (2026). Three-dimensional vision-based recognition of guitar chords. *Computer Music Journal*. https://doi.org/10.1162/COMJ.a.690

Mauch, M., & Dixon, S. (2014). pYIN: A fundamental frequency estimator using probabilistic threshold distributions. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 659–663). https://doi.org/10.1109/ICASSP.2014.6853678

McFee, B., Raffel, C., Liang, D., Ellis, D. P. W., McVicar, M., Battenberg, E., & Nieto, O. (2015). librosa: Audio and music signal analysis in Python. In *Proceedings of the 14th Python in Science Conference* (pp. 18–24). https://doi.org/10.25080/Majora-7b98e3ed-003

Mohammadian, R., Pourroostaei Ardakani, S., & Zaraii Zavaraki, E. (2026). A systematic review of technology-enhanced music learning models: Impacts on music education. *Computers and Education Open, 10*, 100354. https://doi.org/10.1016/j.caeo.2026.100354

Ng, K., & Nesi, P. (2008). i-Maestro: Technology-enhanced learning and teaching for music. In *Proceedings of the International Conference on New Interfaces for Musical Expression* (pp. 225–228). https://doi.org/10.5281/zenodo.1179605

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, n71. https://doi.org/10.1136/bmj.n71

Pati, K. A., Gururani, S., & Lerch, A. (2018). Assessment of student music performances using deep neural networks. *Applied Sciences, 8*(4), 507. https://doi.org/10.3390/app8040507

Percival, G., Wang, Y., & Tzanetakis, G. (2007). Effective use of multimedia for computer-assisted musical instrument tutoring. In *Proceedings of the International Workshop on Educational Multimedia and Multimedia Education (EMME '07)*. ACM.

Provenzale, C., Di Tommaso, F., Di Stefano, N., Formica, D., & Taffoni, F. (2024). Real-time visual feedback based on MIMUs technology reduces bowing errors in beginner violin students. *Sensors, 24*(12), 3961. https://doi.org/10.3390/s24123961

Raffel, C., McFee, B., Humphrey, E. J., Salamon, J., Nieto, O., Liang, D., & Ellis, D. P. W. (2014). mir_eval: A transparent implementation of common MIR metrics. In *Proceedings of the 15th International Society for Music Information Retrieval Conference* (pp. 367–372).

Riley, X., Edwards, D., & Dixon, S. (2024). *High resolution guitar transcription via domain adaptation* (arXiv:2402.15258). arXiv. https://doi.org/10.48550/arXiv.2402.15258

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153–189. https://doi.org/10.3102/0034654307313795

Tavares, T. F., Barbedo, J. G. A., Attux, R., & Lopes, A. (2013). Survey on automatic transcription of music. *Journal of the Brazilian Computer Society, 19*, 589–604. https://doi.org/10.1007/s13173-013-0118-6

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist, 46*(4), 197–221. https://doi.org/10.1080/00461520.2011.611369

Wiggins, A., & Kim, Y. E. (2019). Guitar tablature estimation with a convolutional neural network. In *Proceedings of the 20th International Society for Music Information Retrieval Conference*, Delft, The Netherlands.

Wulf, G., & Shea, C. H. (2002). Principles derived from the study of simple skills do not generalize to complex skill learning. *Psychonomic Bulletin & Review, 9*(2), 185–211. https://doi.org/10.3758/BF03196276

Xi, Q., Bittner, R. M., Pauwels, J., Ye, X., & Bello, J. P. (2018). GuitarSet: A dataset for guitar transcription. In *Proceedings of the 19th International Society for Music Information Retrieval Conference* (pp. 453–460).

 © 2026 VT International Press. All rights reserved. E-ISSN: 3139-7026

Xia, G., Chin, D., Zhang, Y., Zhang, T., & Zhao, J. (2020). Interactive Rainbow Score: A visual-centered multimodal flute tutoring system. In *Proceedings of the International Conference on New Interfaces for Musical Expression* (pp. 208–213). https://doi.org/10.5281/zenodo.4813324

Zhang, F., Bazarevsky, V., Vakunov, A., Tkachenka, A., Sung, G., Chang, C.-L., & Grundmann, M. (2020). *MediaPipe Hands: On-device real-time hand tracking* (arXiv:2006.10214). arXiv. https://doi.org/10.48550/arXiv.2006.10214

Zhang, Y., Beh, W. F., Zhang, C., & Pi, S. (2024). Transforming music education through artificial intelligence: A systematic literature review on enhancing music teaching and learning. *International Journal of Interactive Mobile Technologies, 18*(18), 76–93. https://doi.org/10.3991/ijim.v18i18.50545

 © 2026 VT International Press. All rights reserved. E-ISSN: 3139-7026