Computers and Education: Artificial Intelligence 4 (2023) 100125

Contents lists available at ScienceDirect

Computers and Education: Artificial Intelligence

journal homepage: www.sciencedirect.com/journal/computers-and-education-artificial-intelligence

A systematic review of intelligent tutoring systems based on Gross body movement detected using computer vision

T.S. Ashwin *, Vijay Prakash, Ramkumar Rajendran

IIT Bombay, Mumbai, India

ARTICLE INFO

Keywords:
Intelligent tutoring system
Gross body movements
Systematic review
Computer vision
Hand gestures
Body postures
Machine learning
Deep learning

ABSTRACT

The computer vision applications in intelligent tutoring systems (ITS) have enabled its use in various domains such as dance and sports. The adaptation in the tutoring system is based on the analysis of hand gestures, body postures, and movement. All these are termed as gross body movements (GBM). The use of these in the intelligent tutoring systems is termed as GBM-ITS in this article. There is no survey paper that considers the use of GBM-ITS in different domains. A systematic process is followed to address six review questions (RQs) by considering the 33 articles published between 2010 to 2022. A brief discussion regarding the methods adopted for analysis and the performance metrics used for evaluation has been made. The use of devices for the purpose of the study is analyzed with the need for its use. The feedback mechanisms adopted by the studies are reviewed. The information derived from the survey by addressing the RQs has been summarised as observations. This review results indicate the impact of GBM-ITSs in various domains and its potential to reduce human interventions significantly in the near future. Some of the future directions the review results indicated are: 1. the state-of-the-art computer vision methods for detection and tracking along with the temporal aspects are not much explored in GBM-ITS. 2. the current GBM-ITS are designed for beginners, and there is a vast scope to extend it for intermediate and experts. 3. the test users are considerably less, and there is a need for large-scale implementations and testing for generalizability or to check robustness. 4. more importance can be given to privacy and ethics as the GBM-ITS is performing better 5. since most methods use machine/deep learning methods, dataset dissemination will increase the design and development of GBM-ITS. 6. mechanism used in providing feedback needs to be evaluated and optimized.

1. Introduction

Intelligent tutoring systems (ITS) provide personalised content based on learners' performance and preferences. ITS provides feedback to the learners usually without interference of a human tutor (Freedman et al., 2000). In recent trends, the adaptive content provided by the ITS is based on the learners' cognitive and affective states (Rajendran et al., 2018, Munshi et al., 2018). The adaptive logic required for the same has been developed using the machine learning methods (Aranha et al., 2019). During 1960, before the introduction of the term ITS, there was intelligent computer assisted instruction (ICAI) where computers were used only to provide instructions (Carbonell, 1970). These ICAIs used only the interaction log data to model/understand the learners' performance and to provide personalised instructions. The recent advancements in ITS considers not only the interaction data but also the learners facial expressions, body postures and hand gestures along with the data from various physiological sensors (like electroencephalography (EEG) and electrocardiography (ECG)) to provide adaptive content (D'mello & Kory, 2015). Moreover, the content in ITS can be delivered via devices other than computers such as virtual reality (VR), mobile, kinect (Olney et al., 2012, D'Mello et al., 2007).

Approximately two decades ago, due to lack of technological advancements it was not convenient to observe or use the body movements and other aspects in the tutoring systems (Mousavinasab et al., 2021). The wide range of development in the technology has equipped the researchers to collect data regarding all these movements for the purpose of their study which has led to the defining of such movements as gross body movements (GBMs). These refer to the hand gestures, body postures, head movements and the movements of the large muscles of the arms, legs, and torso (Lara & Labrador, 2012, Wang et al., 2019, D'Mello & Graesser, 2009, Bosch et al., 2015). Generally, these GBMs can be easily analysed using computer vision (CV) methods. The computer vision methods are a part of artificial intelligence facilitating the systems to extract the required information from any image frame,

* Corresponding author.
E-mail addresses: ashwindixit9@gmail.com (T.S. Ashwin), prakashvijay649@gmail.com (V. Prakash), ramkumar.rajendran@iitb.ac.in (R. Rajendran).

https://doi.org/10.1016/j.caeai.2023.100125
Received 30 June 2022; Received in revised form 30 November 2022; Accepted 9 January 2023
Available online 13 January 2023

video content or other visual inputs. These inputs are used to take actions or make recommendations (Voulodimos et al., 2018). Recently, the GBMs of learners were inculcated into a few tutoring systems in domains where there was an absolute necessity to study the learners hand gestures or postures to analyse their performance as in case of sports or dance. We call such ITSs as gross body movement based intelligent tutoring system (GBM-ITS). Most of the tutoring systems are designed under an E-learning environment where the learners sit in front of the device like a computer. In such circumstances, the body movements will not be of much importance. Other than these, in other domains like co-curricular and extracurricular activities, the intelligent tutoring systems will consider the gesture, timing, positions, the entire body movement tracking and so on through the computer vision. As these detected movements were used as inputs for the intelligent tutors, the generalized term was given as gross body movements-intelligent tutoring system. This can be simply mentioned as the intelligent tutoring systems which are based on the gross body movements detected through computer vision.

In GBM-ITS detecting the step by step movements to arrive at a single stance is challenging and important. For instance in sports or dance, these movements are required to be analysed effectively and efficiently to measure the learners performance. This challenge can be achieved through the use of deep learning architectures which consider temporal analysis. Deep learning is a subset of machine learning which is under artificial intelligence. This emulates the ways humans obtain knowledge. The need to analyse spatial or temporal data in a computer vision to make it more achievable through the implementation of deep learning architectures in it (Shrestha & Mahmood, 2019). The recent advancements in computer vision and deep learning have led to the development of effective methodologies that can perform spatio-temporal analysis of human pose, gesture, posture, and speech recognition in real-time. These methodologies can be adopted in the various domains of intelligent tutoring systems such as sports, dance, musical instruments, and others where the GBMs are detected, to understand and analyse the learner's performance (D'mello & Kory, 2015, Shrestha & Mahmood, 2019, Rajšp & Fister, 2020).

The survey presented in this study provides the details of a few domains in which GBM-ITS are designed, the methodologies used, the ways in which feedback is provided and the state-of-the-art in GBM-ITS. To the best of our knowledge, this is the first survey paper to consider these domains and the use of GBM-ITS in it. The research questions (RQs) addressed in this study are listed below. These RQs are more targeted toward the computer vision developers or designers of GBM-ITSs. The RQs and the corresponding details provide the direct and basic information required by anyone from the computer vision field. It starts with analyzing the state-of-the-art method usage with its respective performance metrics that can be considered, the type of data used if classification is done, then the class labels, the size of data, and complete dataset details. And finally, way of using the entire methodology to provide a suitable type of feedback.

- RQ1: What are the various domains in ITS based on body movement applications detected using computer vision?
- RQ2: What are the methods used in GBM-ITSs to detect GBMs?
- RQ3: What are the performance metrics used in the GBM-ITS?
- RQ4: What are the devices used in the GBM-ITS?
- RQ5: Who are the target participants of the GBM-ITS and what are their proficiency level?
- RQ6: What type of feedback is provided by the GBM-ITS?

The rest of the paper is described as follows. Section 2 contains the information on the article selection process. Section 3 discusses the addressed research questions in detail. Section 4 contains details of observations and suggestions. The final section concludes the review article with possible future directions.

Table 1
Search Query Details.

Base Boolean used in the Search Query	(tutor* OR coach*) AND (image* OR video*) AND (gesture OR posture OR head OR body OR hand)
Databases	ACM, IEEE, Springer, Science Direct, Taylor and Francis, Scopus, and Google Scholar
Period	Jan 2010 to May 2022
Article Types	Journal papers, Conferences, and Book Chapters
Language	English

2. Materials and methods

The articles selected in this review follow a systematic process PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) (Moher et al., 2009, Page et al., 2021). The review process consists of 4 main steps: identification, screening, selection, and inclusion.

2.1. Identification

The current study addresses the research questions mentioned in the introduction section 1 from the selected papers. The first step is to find the relevant keywords and search databases.

We followed the exploratory analysis using google scholar to find the keywords. Once the keywords are shortlisted, these are used to search in the following databases namely; 1. ACM[1] 2. IEEE,[2] 3. Springer,[3] 4. Science Direct,[4] 5. Taylor and Francis[5] 6. Scopus[6] and 7. Google Scholar.[7]

The keywords used in the study are tutor, coach, train, system, learn, student, intelligent, gesture, posture, face, head, image, video, multi, and vision. We used loose/approximate phrases for all these keywords; for example, tutor* was given as a search keyword that maps to tutoring, tutor, and so on. All these keywords are searched in the title, abstract and keywords. Few keywords like pose, hand, fingers, wave, body, and ITS are searched in the contents of the entire document. The details of the exact boolean used in the search query are mentioned in Table 1. Keywords such as multi*, vision, intelligent, learn*, and student* are also used along with the base boolean to obtain more related articles. A total of 1729 articles are obtained after the initial search (Scopus: 1110, IEEE: 169, ACM: 101, T&F: 81, Springer: 17, Elsevier: 93, Scholar: 137. Additional searches of 21 articles are from other databases like MDPI and Inderscience) as shown in Fig. 1.

2.2. Screening

The inclusion (I) and exclusion (E) criteria are adopted in the screening step based on the title, abstract, and keywords.

- I1: Articles published between Jan 2010 to May 2022
- I2: Articles that are published and fully available on scientific databases
- I3: Articles from journal papers, conferences, and book chapters
- I4: Articles that use gross body movements and computer vision to analyze the data and accordingly provide the feedback in the tutoring system

- E1: Outside the period of coverage

[1] https://www.acm.org/.
[2] https://ieeexplore.ieee.org/.
[3] https://www.springer.com/.
[4] https://www.sciencedirect.com/.
[5] https://taylorandfrancis.com/.
[6] https://www.scopus.com/.
[7] https://scholar.google.com/.

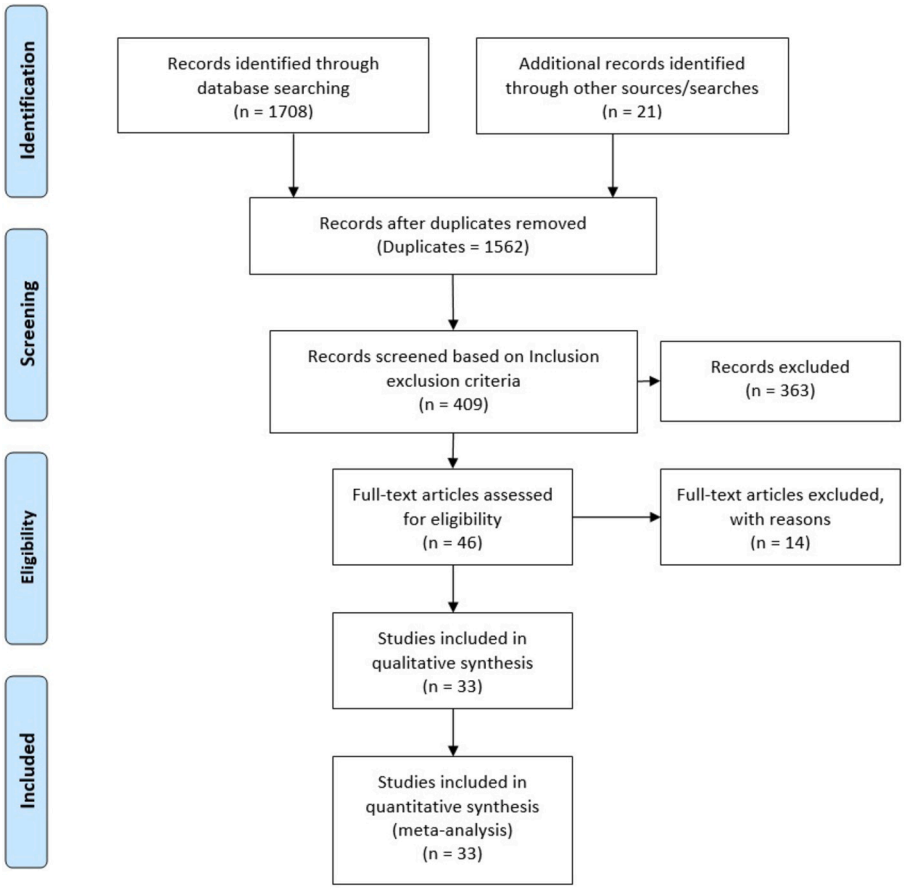

Fig. 1. Flow of the systematic review process.

- E2: Non-English articles
- E3: Conceptual papers and unimplemented frameworks.
- E4: Posters presentations and abstract submissions
- E5: Earlier versions of the considered articles.
- E6: Sign language based tutoring systems
- E7: Those articles which were not available in full version

Though there are plenty of articles on the use of various intelligent methods to predict the different aspects of gross body movements, a minimal number of articles discuss a developed tutoring system that uses gross body movement. Hence, we had to consider all these articles irrespective of conference or book chapters.

Sign language detection uses computer vision techniques, but they are excluded in the current review process as several recent review papers exist (Wadhawan & Kumar, 2021, Rastgoo et al., 2021, Papastratis et al., 2021, Sharma & Anand, 2021). From a total of 1729 articles, we removed duplicates[8] and obtained 1562 articles. For these 1562 articles, we applied the inclusion, exclusion criteria and obtained a total of 409 articles.

2.3. Eligibility

The 409 articles were classified into three categories, the first category of papers are the selected papers (19 papers), the second category needs detailed reading to shortlist (27 papers), and the last category is rejected papers (remaining 363). This categorisation was based on abstract level reading. Subsequently, the first and second category papers are considered in the extraction step (19 + 27 = 46 papers), where two reviewers read in detail and shortlist the second category of papers. After full reading of the 27 papers and repeating inclusion exclusion criteria, 14 papers were selected. The reliability of two reviews for selecting these articles was Cohen's $kappa = 0.93$. This resulted in a total of 33 (19+ 14) papers for the final step.

2.4. Included

Finally, 33 papers were shortlisted for qualitative synthesis and meta-analysis. The flow of the systematic review process is shown in Fig. 1.

3. Discussion and inferences

The annual distribution of the shortlisted 33 articles is shown in Fig. 2. All the RQs are discussed in detail in this section, and corresponding inferences are provided. In the first RQ, the following questions are addressed. If we can classify the selected papers into domains and subdomains, if yes, the related methodology, user, feedback mechanism, and devices used can be cherry-picked for that particular sub-domain while designing and using the GBM-ITS. For every domain, are there any subdomains? For instance, domains like dance have wide styles that significantly vary from each other and are classified into subdomains based on culture and region. Are there any names given to these tutoring systems? If yes, avoid reusing the tutor's name for their new tutor or referring directly to the details of that particular tutor.

RQ 1: What are the various domains in ITS based on gross body movement applications detected using computer vision?

[8] https://www.mendeley.com/. Mendeley Desktop is used to remove the duplicates. All the articles selected in the previous step are loaded using Mendeley web importer and used the .bibtex format. Then "Check for Duplicates" function in the tools is used to find and remove the duplicates.

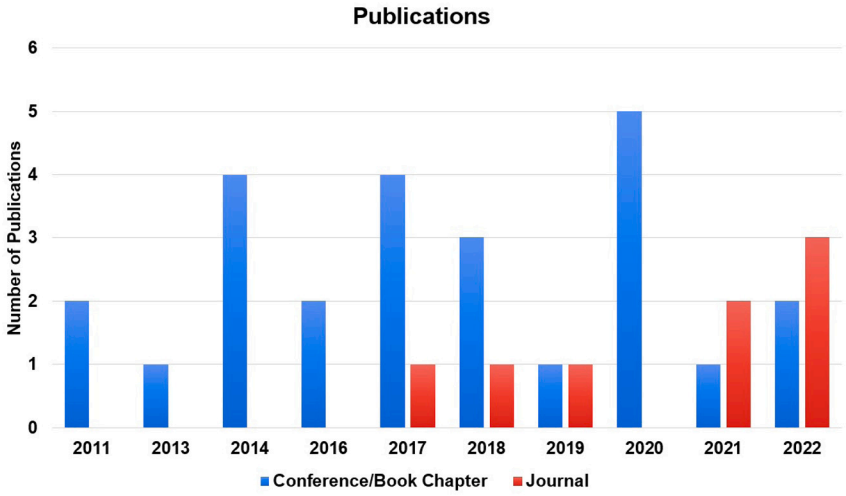

Fig. 2. Annual distribution of the selected articles.

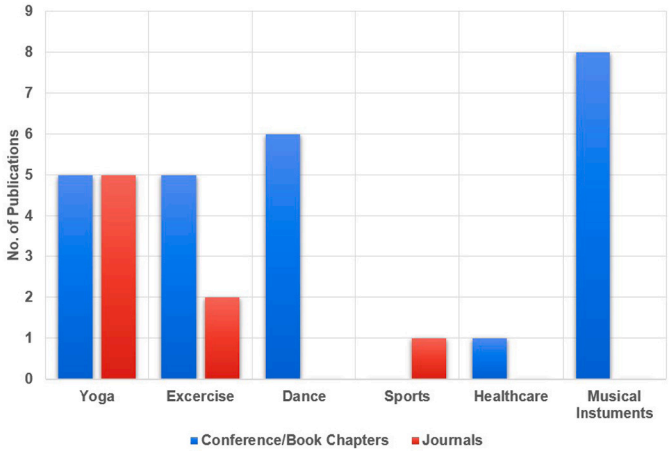

Fig. 3. Distribution of published articles based on domain.

The selected articles belong to five domains such as dance (Aich et al., 2017) (Muangmoon et al., 2017) (Liu et al., 2017) (Ramadijanti et al., 2016) (Majumdar et al., 2014) (Ros et al., 2014), exercise (Görer et al., 2017) (Chen et al., 2013) (Joseph et al., 2022) (Masala & Angdresey, 2017) (Xu et al., 2019) (Movva et al., 2022) (Dittakavi et al., 2022), musical instruments (Molloy et al., 2019) (Huang et al., 2011) (Sun & Chiang, 2018) (Carvalho et al., 2020) (Ritschel et al., 2020) (Johnson et al., 2016) (Rho et al., 2014) (Rigby et al., 2020), yoga (Kale et al., 2021) (Chen et al., 2014) (Patil et al., 2011) (Wu et al., 2021) (Chen et al., 2018) (Rishan et al., 2020) (Anand Thoutam et al., 2022) (Trejo & Yuan, 2018) (Long et al., 2022) (Chaudhari et al., 2021), heath care (Di Mitri et al., 2020) and sports (Mat Sanusi et al., 2021). The distribution of published articles based on the domain is shown in Fig. 3.

In the considered articles, each domain can be divided into subdomains. For example, the dance domain can have subdomains like Bharatnatyam, Thai, Remo, traditional Indonesian dance, and so on, corresponding to cultural and regional dance styles. Similarly, the yoga domain has Ashtanga and Vinyasa yoga. The sports domain has table tennis, and the musical instruments domain has pianos. Some GBM-ITS has names such as YogaST Yoga Tutor (Chen et al., 2014), Infinity Yoga Tutor (Rishan et al., 2020), YOG-GURU (Chaudhari et al., 2021), HasTA (Muangmoon et al., 2017), CPR Tutor (Di Mitri et al., 2020), Table Tennis Tutor (T3) (Mat Sanusi et al., 2021), Mixed Reality Piano Tutor (Molloy et al., 2019), PETA (Carvalho et al., 2020), and piARno (Rigby et al., 2020). Nao is a robot, and it is used in the GBM-ITS (Ros et al., 2014).

As there are several GBM-ITSs available, in the coming RQs, we will discuss the methods used and their performance in detail to understand the impact of these systems.

RQ 2: What are the methods used in GBM-ITSs to detect GBMs?

Table 2 explains the computer vision methods considered in the GBM-ITS along with the performance metrics considered for evaluation of the same.

Generally most of these studies consider video streams or image frames. These image frames are used for classification detection or recognition. The studies related to a period of 2010 to 2016, considered feature based computation methods existing prior to deep learning like skeletal tracking method, histogram of normal vectors, contour based methods. Post this period, use of machine learning methods and libraries like Adaboost, OpenCV and others. As most of the domains under consideration are based on temporal data as to how they arrived at a particular stance. As the LSTM (long short-term memory), RCNN (region-based convolutional neural network), reinforcement learning and many other deep learning methods analyse the temporal data efficiently in other domains (Shrestha & Mahmood, 2019), most recent studies considered under review significantly use the deep learning methods as well.

Initial methods would consider either body contour or skeletal joint positions and define a centroid to analyse. Based on the deviation from this they arrive at the match or mismatch regarding the postures. Techniques like Adaboost were used to improve the robustness of classification. As there was a need for the consideration of posture, movement as well as speed, the temporal data analysis was necessary. This led to the use of deep learning architectures. Almost all the methods are used for classification and are supervised learning.

RQ 3: What are the performance metrics used in the ITS using GBM?

The performance of the classification methods used in these intelligent tutoring systems is generally measured using accuracy (17 out of 33 articles). As complete classification information may not be arrived at from mere accuracy few studies have considered precision, recall, F1 score and confusion metrics as well (11 out of 17 articles). In absence of classification, few studies have defined their results through curves like RMSE (Root Mean Square Error) curve, DTW curve, *p*-value and others (Table 2). Few studies which have considered devices or libraries like Kinect or OpenCV have not emphasised on the accuracy and other metrics, rather they have either collected the participants feedback through questionnaires to analyse their performance or have totally relied upon the Kinect sensors. Few other studies have taken the feedback from a domain expert to analyse the performance of it. Generally the ques-

Table 2
Computer vision methods considered in the GBM-ITS along with the performance metrics.

Author, Year	Method	Performance in %	Performance Metrics
(Patil et al., 2011)	SURF (Speeded Up Robust Feature) algorithm	–	–
(Huang et al., 2011)	Geometrical feature gaining translational and rotation matrix	Average Error 1.97	–
(Chen et al., 2013)	Contour and skeleton computation	Acc: 98.67	Accuracy and CM
(Chen et al., 2014)	Star Skeleton	Acc: 99.33	Accuracy and CM
(Majumdar et al., 2014)	Kinect API (Kinect Skeletal Tracking algorithm)	Manual Verification by expert	–
(Ros et al., 2014)	-	Observational analysis of the recordings	–
(Rho et al., 2014)	Voting algorithm	Acc: 94.80	Accuracy
(Ramadijanti et al., 2016)	Skeletal Tracking method	Feedback as measure	–
(Johnson et al., 2016)	Histograms of Normal Vectors (HONV)	Acc: 96.0	Accuracy
(Görer et al., 2017)	COBYLA (Constrained optimization by linear approximation) algorithm	Mean error from 0.32 to .10	Mean error
(Masala & Angdresey, 2017)	DTW	–	–
(Dittakavi et al., 2022)	DenseNet	Acc: 81, 79, & 62 (3 Databases)	Accuracy, Approval Rate, *p*-value
(Muangmoon et al., 2017)	Kinect API Optical motion capture system	–	–
(Liu et al., 2017)	Three cosine eigenvalues	Feedback	–
(Chen et al., 2018)	Contour and Skeleton Computation	Acc: 76.22 to 99.51	Accuracy
(Trejo & Yuan, 2018)	Adaboost	Acc 94.78	Accuracy and CM
(Aich et al., 2017)	Dynamic Time Warping (DTW)	DTW Curve	DTW Curve
(Sun & Chiang, 2018)	GMM	Error rate upto 16%	–
(Xu et al., 2019)	HMM	RMSE Curve	RMSE Curve
(Molloy et al., 2019)	ArUco AR (Augmented Reality) Markers OpenCV	SUS: 82	SUS
(Rishan et al., 2020)	OpenPose and MaskRCNN	Acc: 99.91	Accuracy and CM
(Di Mitri et al., 2020)	LSTM	max Acc .98	Accuracy, Precision, Recall, F1
(Carvalho et al., 2020)	MD-DTW (Multidimensional Dynamic Time Warping) algorithm	Likert Scale	–
(Ritschel et al., 2020)	Reinforcement Learning	–	–
(Rigby et al., 2020)	MusicXML	Liker Analysis 1.91 to 4.23 And SUS 74.80	Accuracy and SUS
(Kale et al., 2021)	Two-layer hierarchical Model	Acc: 98.29	Accuracy
(Chaudhari et al., 2021)	CNN (Convolutional Neural Network)	Acc: 95.00	Accuracy, Precision, Recall, F1
(Mat Sanusi et al., 2021)	LSTM	Acc: 0.62	Accuracy, Precision, Recall
(Wu et al., 2021)	BlazePose	Acc: 83.21	Accuracy, Precision, Recall, F1
(Anand Thoutam et al., 2022)	MLP (Multilayer perceptron)	Acc: 99.58	Accuracy
(Long et al., 2022)	TL-MobileNet-DA	Acc 98.43	Accuracy, CM Sensitivity and Specificity
(Joseph et al., 2022)	PoseNet	Acc: 98.32	Accuracy
(Movva et al., 2022)	PoseNet	Acc: 96.77	Accuracy and CM

Acc: Accuracy; CM: Confusion Matrix; SUS: System Usability Scale.

tionnaires are either general, user experience questionnaires or system usability scale (SUS).

RQs 2 and 3 provided the methods and performance metrics details. In the next RQ, we discuss the devices used in these applications as the methods and the performance are related to the devices or sensors used.

RQ 4: What are the devices used in the GBM-ITS?

Table 3 explains the devices used by each of these studies along with the consideration of hand gestures and body postures.

Out of the domains under consideration, most of them require the study of entire body postures as in case of dance or yoga, but for few others, mere hand gestures are considered as in case of playing piano, dance hand gestures and others. In a sport like table tennis, though the entire body movement needs to be analysed, the considered study emphasises on fore and back hand movements alone. In order to implement their method in the tutoring system, authors have used devices such as camera (2D, RGB), depth camera, web camera, mobile cameras and head mounted devices like HTC Vive. As Kinect is the most effective and popular device among the depth cameras, it is preferred and used significantly in most of the studies. In studies considering the mobile phones, inbuilt sensors such as an accelerometer and a gyroscope capable of capturing the participant's motion data were also used to make their method more robust by including multimodality.

RQ 5: Who are the target users of the GBM-ITS and what are their proficiency levels?

Table 4 is a brief about the target users of the GBM-ITS and their level of proficiency in the domain along with a few other details. There is also a discussion about the involvement of an expert for evaluation of the methods or the overall performance of the ITS. The consideration of the ground truth or annotations for comparison as briefed.

Table 3
Devices used along with the hand gesture and body posture details.

Devices Used	HG	BP	Articles
Camera	n	y	(Patil et al., 2011), (Dittakavi et al., 2022), (Molloy et al., 2019), (Chaudhari et al., 2021), (Wu et al., 2021), (Anand Thoutam et al., 2022)
Camera	y	n	(Huang et al., 2011), (Molloy et al., 2019), (Rigby et al., 2020)
Kinect	n	y	(Chen et al., 2013), (Chen et al., 2014), (Ramadijanti et al., 2016), (Görer et al., 2017), (Masala & Angdresey, 2017), (Muangmoon et al., 2017), (Liu et al., 2017), (Chen et al., 2018), (Kale et al., 2021), (Di Mitri et al., 2020), (Xu et al., 2019), (Aich et al., 2017), (Trejo & Yuan, 2018), (Joseph et al., 2022)
Kinect	y	n	(Majumdar et al., 2014), (Johnson et al., 2016), (Di Mitri et al., 2020), (Mat Sanusi et al., 2021)
Depth Camera	y	n	(Rho et al., 2014)
Intel Realsense	y	n	(Johnson et al., 2016)
Robot	n	y	(Görer et al., 2017)
Mobile	n	y	(Dittakavi et al., 2022), (Rishan et al., 2020), (Movva et al., 2022)
Mobile	y	n	(Sun & Chiang, 2018), (Mat Sanusi et al., 2021)
Webcam	y	n	(Sun & Chiang, 2018), (Carvalho et al., 2020)
HTC Vive	y	y	(Molloy et al., 2019)
Web Camera	n	y	(Rishan et al., 2020), (Long et al., 2022)
Myo	y	y	(Di Mitri et al., 2020)
VR Headset (HTC Vive Pro)	y	n	(Rigby et al., 2020)

HG: Hand gesture; BP: Body Posture; n: No; y: Yes.

The tutoring systems are designed mostly for beginners except in a few cases where even the intermediate and experts in the domain used the tutoring system to enhance their skills. Since the methods used are either classification or pattern matching algorithms, these methods need to be backed up by a ground truth. The ground truth is considered as follows: an average of several participants' poses, majority voting method, data collected from an expert, manual verification or observation from an expert, use of annotation tools like visual inspection tool to annotate the dataset, and the annotated data is considered the ground truth. In musical instruments like piano, the electronic signals generated from the electronic piano are considered the ground truth. In a few studies, even after considering the ground truth, expert advice regarding the method and the tutoring system was obtained. All the studies involved human subjects, but only a few studies have explicitly mentioned the ethical statement or consent from the participants in their article.

To understand the methodology considered to the target users in the selected articles better, Table 5 explains the data size, and the corresponding information about the collected data.

It is inferred from this table that the sample size of subjects considered for the study varies between 1 to 45. (Liu et al., 2017) tested their tutoring system on Tujia Museum staff but did not mention the number of participants. The data used in the considered studies are video streams or image frames. High quality (ranges around 1280 × 780), medium quality (ranges around 640 × 480), and (ranges around low-quality 320 × 200) videos are considered in these studies. Some studies explicitly used all three to increase the robustness of the pose detection methods. The most commonly used frame rates are 30 and 15 frames per second.

Most studies considered collecting data multiple times for each class label/category in the data collection process and were balanced and imbalanced datasets. The participants are of different ages ranging from 6 to 88 years. The ratio of men and women participants was also provided in some cases (Long et al., 2022), (Görer et al., 2017), (Xu et al., 2019), (Movva et al., 2022), (Muangmoon et al., 2017), (Majumdar et al., 2014), (Carvalho et al., 2020).

Table 4
Target users of the GBM-ITS.

Author, Year	TU	Eu	Annotation/Ground Truth
(Kale et al., 2021)	B	y	-
(Chen et al., 2014)	B	-	Majority voting
(Patil et al., 2011)	B	-	-
(Wu et al., 2021)	B	-	-
(Chen et al., 2018)	B	y	An yoga expert is asked to judge
(Rishan et al., 2020)	-	-	Used an existing dataset
(Anand Thoutam et al., 2022)	-	-	Average values or angles are calculated for every pose by considering all poses done by everyone
(Trejo & Yuan, 2018)	B	y	-
(Long et al., 2022)	-	-	-
(Chaudhari et al., 2021)	-	-	-
(Görer et al., 2017)	-	-	-
(Chen et al., 2013)	B	y	Yoga expert is asked to judge
(Joseph et al., 2022)	-	-	-
(Masala & Angdresey, 2017)	B	y	Action data of the real trainer
(Xu et al., 2019)	B	y	Expert assessor is added to evaluate the learning of each participant.
(Movva et al., 2022)	B	y	Data from internet
(Dittakavi et al., 2022)	-	y	Used an existing dataset and experts also
(Aich et al., 2017)	B	y	Recorded data from the expert
(Muangmoon et al., 2017)	B	y	Recorded data from the expert
(Liu et al., 2017)	B	y	Professional hand-waving choreographer
(Ramadijanti et al., 2016)	B	-	-
(Majumdar et al., 2014)	B	y	Designed by experts
(Ros et al., 2014)	B	y	Trained robot
(Di Mitri et al., 2020)	B	y	Visual inspection tool
(Mat Sanusi et al., 2021)	B	y	Visual inspection tool
(Molloy et al., 2019)	B	-	-
(Huang et al., 2011)	-	-	-
(Sun & Chiang, 2018)	B	-	-
(Carvalho et al., 2020)	B, I, E	-	-
(Ritschel et al., 2020)	-	y	Trained robot
(Johnson et al., 2016)	I, E	y	Teachers are used
(Rho et al., 2014)	-	-	The electronic signals of the electronic piano is the ground truth
(Rigby et al., 2020)	B	-	Electronic notes

A: Article; TU: Target User; B: Beginner; I: Intermediate; E: Expert; Eu: Experts used in the study; y: Yes.

The use devices, method, target user and the feedback mechanism is directly dependent on the considered class labels. Table 6 explains the details of the number of classes and the class labels given to the poses in each of the considered articles.

The number of class labels used according to the considered data, range from 1 to 82 (Yoga 82 - is a dataset with 82 class labels). In a few domains instead of classifications, scoring systems have been adopted wherein a learner will be graded from 0-100 based on the accuracy of representing the concepts in order to proceed to the next level of learning.

In the previous RQs, we discussed the methodology, performance, devices, and data used along with the target users. In the next RQ, we will discuss the feedback mechanism used by each GBM-ITS.

RQ 6: What type of feedback is provided by the GBM-ITS?

A device can communicate to the users through various modalities such as vision (screen outputs), audition – (audio outputs), tactition – (vibrations or other movements), gustation (taste), olfaction (smell), thermoception (heat), nociception (pain) and equilibrioception (balance). Vision and audition are the most commonly used modalities in the intelligent tutoring systems. The feedback information can be represented through text, numerical data, visual, speech or audio, and others.

Table 5
Collected data details.

Article	Data Size	Data Details
(Kale et al., 2021)	25	1750 video sequences of 7 Yogasanas
(Chen et al., 2014)	5	25 video clips for each posture, and totally 300 clips to be recognized
(Patil et al., 2011)	2	Reference video and actual/practitioner video data
(Wu et al., 2021)	45	2 Datasets, 45 categories and 1931 images are selected and 3000 triplet examples of high-quality, medium-quality, and low-quality images are considered
(Chen et al., 2018)	5	5 people perform three asanas 5 times with max 8 frames per second
(Rishan et al., 2020)	15	Ten males, five females and all the videos are in 1366 x 768 resolution at 30 frames per second using a common web camera from a computer
(Anand Thoutam et al., 2022)	15	Total videos of the 6 poses are 70, and total instances combining the 6 poses are 350 with 30 fps
(Trejo & Yuan, 2018)	3	Expert yoga trainer, 5 clips were recorded for each Yoga pose with a framerate of 30 frames/sec.
(Long et al., 2022)	8	1,120 images of 14 poses
(Chaudhari et al., 2021)		Yoga-82 dataset
(Görer et al., 2017)	27	Each motion is performed 5 times. Each participant then exercised with the robot for about 10 min. Study is done in different section with different no of the users e.g. 9 (age 25-35), 6 (age 70-80), 12 elderly(age 70-88).
(Chen et al., 2013)	5	Five practitioners perform each of the three asanas five times
(Masala & Angdresey, 2017)	2	Trainer and user data including slow tap dance that focuses on timing and beats
(Xu et al., 2019)	10	There are totally 10 participants in the user study aged between 6 and 15 (average age = 10.7 years),
(Movva et al., 2022)	23	461 images (187 mountain poses, 153 triangle poses, and 121 warrior poses) 23 participants aging between 18–25
(Dittakavi et al., 2022)	15	12 yoga, 1 pilates and 2 kung fu instructors who studied Pose Tutor's outputs
(Aich et al., 2017)	8	Bharatanatyam Adavus data is collected from both experts as well as learners. Tested on 8 learner's video
(Muangmoon et al., 2017)	2	Two captured Thai dances by some volunteers
(Liu et al., 2017)	Many	The staff in Tujia Museum learned the Hand-waving dance in front of the Kinect.
(Ramadijanti et al., 2016)	25	Elementary, junior high and high school students. The number of users tested with the application is 25 people, who do not have experience performing basic dance.
(Majumdar et al., 2014)	36	3 classes of a girl's school where Bharatanatyam is taught as a subject (standards: 8,9,11; duration: 40 minutes each) Each class had 12 girls on an average. We captured photographs of class in progress and also a set of 28 Hastas of 3 random students
(Ros et al., 2014)	11	11 children during 4 days. Each child was expected to interact in three sessions each in different days
(Di Mitri et al., 2020)	10	20 sessions from 10 participants with 2223 CCs (chest compressions)
(Mat Sanusi et al., 2021)	2	One user and one expert. The participant dataset resulted in a total of 33 sessions and 510 recorded strokes
(Molloy et al., 2019)		Most participants were students aged between 18-21, 5 were adults (aged 30-50) and only 6 had played piano previously.
(Huang et al., 2011)		Recognition and tracking took about 30 ms, and the component Pose Estimation took 20 ms. Data is collected at every 15 fps
(Sun & Chiang, 2018)	20	20 subjects who had never learned piano
(Carvalho et al., 2020)	6	Volunteers are piano students, 16.7% and 83.3% are beginner or have intermediary experience in piano playing, 66.6% has experience in wrist movement training while 33.3% have no experience in that technique and 66.6%, 16.7% and 16.7% have beginner, intermediary and advanced experience in any music learning aid software.
(Johnson et al., 2016)	2	P1 is a piano teacher that plays at an advanced level and Pianist P2 plays at an intermediate level.
(Rho et al., 2014)	1	Performer plays music scores with an electronic piano

Most of the selected studies consider visual modality to provide visual and textual feedback. In a few cases, audio feedback is provided using headphones or speakers. Visual feedback includes providing the skeletal joints image of the user in comparison with the actual image (ground truth), representing the percentage of deviation by superimposing the skeletal joints images using different colours. Also, the colour coding is significantly used to describe the poses' mismatch.

Textual feedback is displayed (visual modality) to the user, which can be a detailed description to improve or a one-word message such as "Decent", "Good", "Great", "Impressive," or a short information such as

Table 6
Class labels used in the selected studies.

A	NCL	Class Labels
(Patil et al., 2011)	1	Chakrasana
(Huang et al., 2011)	-	All the keys in the keyboard and the fingers
(Chen et al., 2013)	3	Tree, warrior III, and downward-facing dog
(Chen et al., 2014)	12	(1) Tree, (2) Warrior III, (3) Downward-Facing Dog, (4) Extended Hand-to-Big-Toe, (5) Chair, (6) Full Boat, (7) Warrior II, (8) Warrior I, (9) Cobra, (10) Plank, (11) Side Plank, and (12) Lord of the Dance
(Majumdar et al., 2014)	28	28 Asamyukta hastas
(Ros et al., 2014)	2	Request response pair
(Rho et al., 2014)	2	Tempo and Intensity, Perfect, Good and Miss
(Ramadijanti et al., 2016)	14	Tindak motion, Gendrug Lombo motion, Lawung motion, Sabetan motion, Gendewa motion, Ukel Kurmo motion, Pentangan Kanan Kiri, Buang Sampur motion, Ukel motion, Ukel motion, Tumpang Tali motion, Ayam Alas motion, Nglandak motion,Tanjak Tancep Sembahan motion
(Johnson et al., 2016)	3	(a) Flat Hands (b) Low Wrists (c) Correct
(Görer et al., 2017)	8	Left side lumbar stretching, right side lumbar stretching, back strength, upper arm stretching, swinging arms, side-front arm raise motions. The leg motion set is composed of knee extensions (both knees) and ankle exercise both ankles).
(Masala & Angdresey, 2017)	2	Matched and Unmatched
(Dittakavi et al., 2022)	82, 32 and 7	Yoga-82, Pilates-32, and Kungfu-7 datasets
(Muangmoon et al., 2017)	11	Go In (Come), Go Out (Go), Happy, Love, Sad, Shy, Smile, Walk, Angry, Laugh, Cry
(Liu et al., 2017)	-	Contains several frames of the static move. After scoring 65, they can move to the next one.
(Chen et al., 2018)	12	(1) Tree, (2) Full Boat, (3) Downward-Facing Dog, (4) Extended Hand To Big-Toe, (5) Chair, (6) Warrior I, (7) Warrior II, (8) Warrior III, (9) Cobra, (10) Plank, (11) Side Plank, and (12) Lord of the Dance
(Trejo & Yuan, 2018)	6	Dragon flow. (b) Gate pose. (c) Tiger flow. (d) Tree pose. (e) Triangle pose. (f) Warrior flow
(Aich et al., 2017)	8	Bharatanatyam Adavus
(Sun & Chiang, 2018)	2	Two elementary level tune (28 white keys and 20 black keys and Fingers)
(Xu et al., 2019)	2	Well Learned and Notwell
(Molloy et al., 2019)	-	Score out of 100
(Rishan et al., 2020)	6	Bhujangasana (Cobra Pose), Padmasana (Lotus Pose), Shavasana (Corpse Pose), Tadasana (Mountain Pose), Trikonasana (Triangle Pose) and Vrikshasana (Tree Pose)
(Di Mitri et al., 2020)	5	classRelease, classDepth classRate armsLocked and bodyWeight
(Carvalho et al., 2020)	5	5 different musical excerpts from famous piano works
(Ritschel et al., 2020)	4	Musical pieces are divided into four categories with different degrees of difficulty
(Rigby et al., 2020)	5	Five notes
(Kale et al., 2021)	7	Samasthiti Tadasana, Urdhva Hastasana Tadasana, Virabhadrasana II, Vrukshasan, Urdhva Baddhanguliyasana, Utkatasana, Ardhauttanasana
(Chaudhari et al., 2021)	6	Natarajasana and Trikonasana, and Vrikshasana and Virbhadrasana 1 & 2 and Utkatasana
(Mat Sanusi et al., 2021)	2	correct_stroke and Incorrect
(Wu et al., 2021)	45	45 categories of yoga poses
(Anand Thoutam et al., 2022)	6	Cobra (Bhuj), Tree (Vriksh), Mountain (Tada), Lotus (Padam), Triangle (Trik), and Corpse (Shav).
(Long et al., 2022)	14	Bridge posture, cat-cow posture, child posture, cobra posture, corpse posture, downward-facing dog posture, sitting posture, extended side angle posture, warrior II posture, and warrior I posture
(Movva et al., 2022)	3	Mountain, Triangle, and Warrior-1

A: Article; NCL: Number of class labels

"good one", "keep it up", and so on. In one of the studies, textual feedback information is communicated through bluetooth headphone audio (audition modality). A few robotic tutors demonstrate the right way of the movements as feedback. In virtual reality tutors, the feedback is displayed through head-mounted devices. Generally the provided feedback is in real-time. In some of the articles, though they did not mention it explicitly, it is implied that they are in real-time (denoted with '-' in the Table). A major observation with respect to all of these studies was that, most of them have communicated the feedback to the learners but only three papers refer to the evaluation of such feedback mechanisms. These are done through questionnaires, mainly SUS.

Table 7 is a discussion about the feedback system used by the studies considered for this article. The representations (such as text, audio or visual) of the modalities (vision or audition) have been analysed.

Table 7
Details of feedback used in the selected articles.

Article	Re	Mo	RT	E	Details of Feedback
(Kale et al., 2021)	A,V	a,v	Y	N	Frame sequence, body angle position, measured angle at that position and expected range of correct angle.
(Chen et al., 2014)	V	v	-	N	Video of a yoga expert
(Patil et al., 2011)	T	v	Y	N	Automated pose grading for every single image and an overall grade for a final yoga pose image.
(Wu et al., 2021)	V	v	-	N	Contour and skeletal joint based visualised instruction
(Chen et al., 2018)	T	v	Y	N	Sends text messages that includes posture classification, and accuracy
(Rishan et al., 2020)	T	v	Y	N	Difference in angle is mentioned. Based on the sign of difference, whether to rotate joints in clockwise or anticlockwise direction is given as feedback output
(Anand Thoutam et al., 2022)	T	v	-	N	Detailed improvement instructions along with skeletal images
(Trejo & Yuan, 2018)	T	v	Y	N	Fail message of “Incorrect :(“when the user has not achieved the correct posture along with the image
(Long et al., 2022)	A,V,T	a,v	Y	N	A dashboard is used for audio, image and text feedback
(Chaudhari et al., 2021)	A,V,O	a,v	Y	Y	Robot demonstrates the right way along with the verbal feedback
(Görer et al., 2017)	V	v	-	N	Contour and skeletal joint based visualised instruction
(Chen et al., 2013)	V	v	-	N	Skeletal image with colour indications
(Joseph et al., 2022)	T	v	Y	N	Match score and instructions
(Masala & Angdresey, 2017)	A,V,T	a,v	Y	Y	Supplementary material using text, videos and figures
(Xu et al., 2019)	T	v	Y	Y	Uses text to inform user using short sentences like “Nice Job”
(Movva et al., 2022)	V	v	Y	N	Visualisations on difference in joint angle
(Dittakavi et al., 2022)	T,V	v	Y	N	User interface for text and visual feedback
(Aich et al., 2017)	T,V	v	Y	N	One word message like Good, bad and colour codes.
(Muangmoon et al., 2017)	T,V	v	Y	N	Virtual choreographer and movement score
(Liu et al., 2017)	V	v	-	Y	Contour image along with one word message
(Ramadijanti et al., 2016)	T,V	v	-	Y	Through webpage in the form of perfection renderation
(Majumdar et al., 2014)	V,O	v	-	N	Robot demonstrates the right way
(Ros et al., 2014)	A,V,T	a,v	-	N	Error rate plots and audio messages
(Di Mitri et al., 2020)	A	a	Y	Y	Audio feedback using bluetooth headset
(Mat Sanusi et al., 2021)	T,V	v	Y	N	Real-time visual feedback on the correctness of key presses in the form of glowing particles. Instant feedback on note accuracy and timing. Visual end-of-session feedback
(Molloy et al., 2019)	N	N	N	N	Not implemented
(Huang et al., 2011)	A,V	a,v	Y	N	Audio visual feedback with keys and colour coding
(Sun & Chiang, 2018)	A,V	a,v	Y	N	Score mark from 0 to 100%.
(Carvalho et al., 2020)	A,V,T	a,v	-	N	Hints and messages in text and audio
(Ritschel et al., 2020)	N	N	N	N	Not implemented
(Johnson et al., 2016)	T,V	v	Y	N	One word messages like “Perfect” “Good”, “miss” with colour codes
(Rho et al., 2014)	V	v	Y	N	Colour coding for notations

T: Text; A: Audio; V: Visual; O: Others; v: Vision; a: Audition; t: Tactition; N: No; Y: Yes; RT: Real Time; E: Evaluated

The rendering of feedback in real-time is also discussed along with the evaluation of the feedback mechanism. The brief discussion about how these mechanisms are rendered to the users of the GMB-IST by considering the modalities and representations are explained in the last column of Table 7.

4. Observations and analysis

There are a significant number of studies conducted in various domains and subdomains which were found during the literature review process. Most of them have developed computer vision methods to analyse the data but are yet to implement an ITS. These include basketball, volleyball, cricket, tai chi, running, violin, guitar, drums, ballet, belly dance and others (Rajšp & Fister, 2020, Shrestha & Mahmood, 2019). All these domains emphasise on the use of temporal data significantly more than the static data as these are more dependent on the set of movements performed to achieve a static posture. This could be achieved efficiently through the use of recent deep learning architectures leading to the magnified use of the same. This was clearly evident from our study as well (discussed in Table 2) where we observe the implementation of deep learning architectures over a period of time. The significant contributions of these architectures seem to have come to a spotlight as post its introduction, journal publications of articles were observed to have increased (Fig. 2, the distribution of conference articles or book chapters is 76%, and journal articles are 24%).

Gross body movement detection is another challenge. The human body postures differ significantly and are subjective in nature making it hard to be defined and generalised. For instance, the number skeletal joints considered for pose identification differ from one study to another (ranging between 15-25) based on the data size and class labels considered hence can not be generalised. Another instance was that in a few of the domains, mere movement or pose was not sufficient for the study. An additional aspect of holding the stance was also analysed and the users were graded or given scores based on the extent of maintaining the stance. This adds to the challenges faced by the tutors to even detect and render a feedback as well along with generalising an architecture for the same.

Once a GBM-ITS is developed, testing plays a vital role in understanding the impact of the developed system. Hence, the total number of users or learners considered during the studies infers a lot about the robustness and generalizability of the developed methods. The overall users of GMB-ITS considered in this study range between 1-45 (with

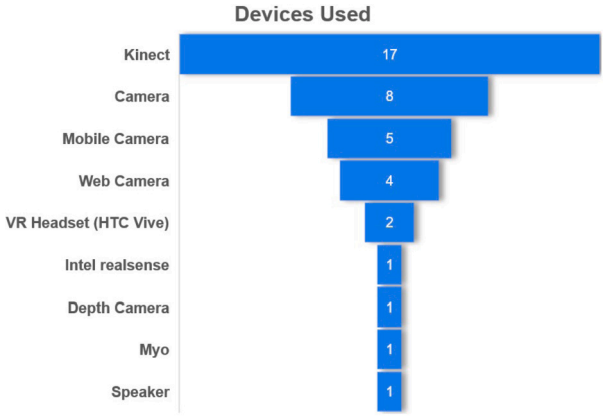

Fig. 4. Devices used in the selected articles.

only 1 study considering 45 users). We used a box plot to understand the distribution of the number of users used in these studies. It is observed that the number of users used in these studies is positively skewed, with a median of 10, mode of 2, and interquartile range of 19 (lower quartile: 3 and upper quartile: 22), indicating a considerably low data size.

Another challenge in computer vision applications is selecting the most appropriate device for data collection or method implementation. A proper selection of devices in the study reduces a lot of effort in hardware requirements. Inorder to detect motion effectively, use of a depth camera will be more feasible. Pose and hand gesture recognition is done with the use of gesture detection and body skeletal detection. Feedbacks can be provided through audio outputs. As all these features are available in Kinect, and can perform efficiently in real-time, it is the most used device among the studies taken into consideration (Fig. 4). As other devices do not have all these features together, those studies considering devices other than kinect, have used additional devices.

It was observed that, in a few papers where the feedback systems were evaluated, the experts have raised a few valuable points which need to be considered in the tutors in order to avoid irritation or demotivation in the users at the beginners levels. This kind of evaluation becomes necessary in every ITS but is hardly considered.

The data used in these studies are all computer vision based. Those which use supervised learning will necessitate data labelling or annotation processes. In order to reuse these data for further research, a few have been mentioned about providing the data on request to the authors and only a few are publicly available. These are mentioned below.

- Yoga-82[9]
- Annotated Bharatnatyam Data Set[10]
- YogaVidCollected[11]

Though the considered studies throw ample light on a number of observations to be made for every single article, those which are significant and general for all the articles are discussed briefly in this section.

4.1. Future directions

The general discussion on the main aspects of computer vision applications are the method, dataset, devices used, and application users are already discussed in the previous subsection. This subsection discusses the challenges observed in the GBM-ITS and the possible future directions from the design and development perspective.

4.1.1. Methodology

Though various methods are used with better performance in the considered articles, several state-of-the-art methods are yet to be explored. The use of the latest technologies may increase the performance of the developed system. For example, multi-instance multi-label learning, faster RCNN, gated recurrent units (GRU), siamese networks, you only look once, and generative adversarial networks are significantly used for human tracking, object recognition, object localization, and identification. GRUs are used for temporal pattern recognition (Shrestha & Mahmood, 2019, Wang et al., 2019). The data can be collected from different sources such as smartphones, CCTV, 2D/3D cameras, or uploaded video data like youtube. Each method can be different based on the data collection mechanism. Selecting a proper method based on the deployment is another challenge, and these are not discussed in the reviewed articles.

Though there are different modalities used in the feedback mechanism, using multimodal data for the analysis is another challenge. For example, understanding the synchrony of dance moves and the audio plays a vital role in the dance field. Now image and audio are the two modalities that should be combined to understand or predict the synchrony. How can these modalities be used together? Can we use a feature or decision-level fusion of these modalities? How to align the image predictions with the audio? How do we aggregate or abstract the results of one modality so that it can be used in a meaningful way for the other modality to predict or classify? These challenges are not addressed in the current studies but can be addressed in future studies related to the methodology.

In real-life applications, data imbalance is another issue that is not addressed in the current study of GBM-ITS studies. All the articles considered classification or regression problems and have different data distributions. For example, in the class labels considered in Bharatanatyam, one class label was more dominantly observed than the other 7 class labels. So in these cases where there is a data imbalance, what type of performance metric needs to be used to understand the performance of the method and how to handle the data imbalance so that the system's performance is not skewed towards one type of data or class label?

The use of methodology or multimodality will be oriented toward the applicability of the same, particularly to a defined problem statement. It may not be feasible to provide common guidelines. Unlike these, data imbalance is quite a common phenomenon while building real-life applications. Using k-fold cross-validation and choosing performance metrics such as precision, recall, F-1 score, Matthews correlation coefficient (MCC), and area under curve (AUC) helps understand the method's performance. There are various methods that address the issue of data imbalance, such as synthetic minority oversampling technique (SMOTE), BalancedBaggingClassifier, and threshold moving (Johnson & Khoshgoftaar, 2019).

4.1.2. Data

The data used in the machine or deep learning-based algorithms are supervised and hence have labelled or annotated data. All the articles that considered these data are not available to the public, and only a few databases are publicly available. And also, the validation of these data is essential, and in some cases, they used experts, and in some cases, it is just majority voting. Considering only one expert or majority voting can have bias, and the method may not be accurate. These issues need to be addressed in the future. How to annotate or validate the data for a given domain? Also, the publicly available data can be used to make the system more robust; hence, training, validating, and testing on different user sets is more important. The availability of more benchmark datasets is of primary importance in this area, and it is one of the future directions.

Managing the team of a large workforce, consistent dataset quality, financial obstacles, and data privacy are some of the significant challenges faced in the annotation process. Some open-source tools, like

[9] https://sites.google.com/view/yoga-82/home.
[10] http://hci.cse.iitkgp.ac.in/.
[11] https://archive.org/download/YogaVidCollected/Yoga_{V}id_{C}ollected.zip.

Label Studio,[12] Computer Vision Annotation Tool[13] (CVAT), and Labelimg,[14] address some of these issues. Also, annotation tools provide an option for semi-automatic annotation, which can help the annotator to speed up the process significantly.

4.1.3. Feedback

It is observed from the literature that the tutoring systems adapted various feedback mechanisms. Only a few have evaluated the feedback system; hence, there is a vast scope in understanding the proper feedback mechanism for each domain based on the target user type. Also, when to provide feedback, in real-time or not, is another challenge, as feedback in the case of sports like table tennis, where giving feedback in real-time or for each frame may not be an ideal solution. If we are providing the feedback at a specific interval or based on the considered class label, then there is a challenge of how to give this feedback. Do we abstract, align and aggregate it in such a way that it is meaningful, and also through what modalities should this feedback be provided?

The type of feedback is subjective mainly to the domain, problem statement, and the devices used. Hence it is up to the discretion of the user.

4.1.4. Ethics and privacy concerns

The articles reviewed did not provide much information on privacy and ethics considered in these GBM-ITS. If we are collecting human data, then how are we discarding it, and if we are using it in a database, then how are we anonymizing this data? A standard method, protocol or guidelines for these can be used to address these issues. How are we maintaining the privacy of the collected data while doing the processing is a more significant challenge in this type of GBM-ITS?

Generally, the training data size will be huge, and to address data privacy, various solutions are available. Few libraries like TensorFlow have dedicated packages like “Tensorflow Privacy (TF Privacy),” which contain multiple functionalities related to privacy that can be used. Also, the developers can adapt different solutions, and a few of them are mentioned below. 1. Secure enclaves are used to execute machine learning workloads in a memory region that is protected from unauthorized access. 2. Homomorphic encryption where machine learning models can be run on encrypted private data using homomorphic encryption, a cryptographic method that allows mathematical operations on data to be carried out on ciphertext instead of on the actual data itself. 3. Secure federated learning, where it builds machine learning models based on data sets that are distributed across multiple devices. With federated learning, multiple data owners can train a model collectively without sharing their private data. 4. Secure multi-party computation where it distributes a large volume of training data among many parties.

4.1.5. Generalization

As the number of target users is not more than 45, the generalizability or the robustness of the GBM-ITS is unclear. Considering data from different age groups or diverse cultural and regional backgrounds will help understand the method and its performance better. Experimenting with a huge number of learners is another challenge, as it includes addressing the devices, methodologies, and databases used and the privacy and ethics related concerns that addresses the global audience.

The generalizing process significantly depends on the data. If the required data is insufficient, then data augmentation (Shorten & Khoshgoftaar, 2019) is one of the solutions as it addresses overfitting. Another solution is invariant risk minimization (Arjovsky et al., 2019) which is a learning paradigm that estimates causal predictors from multiple training environments.

[12] https://labelstud.io/.
[13] https://cvat.org/.
[14] https://github.com/heartexlabs/labelImg.

Apart from the above mentioned future directions, the GBM-ITSs can also incorporate the learner’s engagement. The four types of engagement are emotional, behavioural, cognitive, and agentic (Sinatra et al., 2015). Predicting student engagement and using it as feedback to enhance the learning process in ITS is already there in the literature (Graesser et al., 2004, Ashwin & Guddeti, 2020). GBM-ITS can also include these aspects to improve the feedback mechanism further. Similarly, GBM-ITS can also have the self-regulated learning processes of students, like the affective part, as the image or video data can also be used to recognize students’ engagement and self-regulating learning (Munshi et al., 2018). Since the GBM-ITS already uses image frames, adding the emotional engagement or the affective part will not require additional data collection.

These are some issues and guidelines that need to be considered while designing and developing new GBM-ITS or optimizing the existing GBM-ITS.

5. Conclusion

This study has provided an overview of a few intelligent tutoring systems based on gross body movement detected using computer vision. A systematic review process is adopted to address six review questions. The articles published from Jan 2010 to May 2022 are covered in this study. This review shows that the GBM-ITS is significantly used in domains such as sports, dance, musical instruments, physical exercise, and the healthcare domain. The most commonly used performance metrics were accuracy and confusion matrix. In some cases, feedback from an expert was obtained to evaluate the detection or classification method. The feedback was provided in real-time using text, audio, or visual format in most of the studies. The considered tutoring systems were significantly designed for beginners and only in a couple of them were for intermediate and expert levels. The Kinect and camera were significantly used devices in the GBM-ITS. Devices like Myo, Real sense, and web cameras were also used in a few of the studies for data capturing. In the future, the current computer vision applications based review can be extended to understand the literature on all types of gross body movement related to ITS. A more detailed study can be conducted to understand the challenges faced in each domain. The study can also be conducted to understand the importance of ethics and its use in tutoring systems.

Declaration of competing interest

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

References

Aich, A., Mallick, T., Bhuyan, H. B., Das, P. P., & Majumdar, A. K. (2017). Nrityaguru: A dance tutoring system for Bharatanatyam using kinect. In *National conference on computer vision, pattern recognition, image processing, and graphics* (pp. 481–493). Springer.

Anand Thoutam, V., Srivastava, A., Badal, T., Kumar Mishra, V., Sinha, G., Sakalle, A., Bhardwaj, H., & Raj, M. (2022). Yoga pose estimation and feedback generation using deep learning. *Computational Intelligence and Neuroscience*, *2022*.

Aranha, R. V., Corrêa, C. G., & Nunes, F. L. (2019). Adapting software with affective computing: A systematic review. *IEEE Transactions on Affective Computing*, *12*, 883–899.

Arjovsky, M., Bottou, L., Gulrajani, I., & Lopez-Paz, D. (2019). Invariant risk minimization. Retrieved from arXiv:1907.02893.

Ashwin, T., & Guddeti, R. M. R. (2020). Impact of inquiry interventions on students in e-learning and classroom environments using affective computing framework. *User Modeling and User-Adapted Interaction*, *30*, 759–801.

Bosch, N., D’Mello, S., Baker, R., Ocumpaugh, J., Shute, V., Ventura, M., Wang, L., & Zhao, W. (2015). Automatic detection of learning-centered affective states in the wild. In *Proceedings of the 20th international conference on intelligent user interfaces* (pp. 379–388).

Carbonell, J. R. (1970). Ai in cai: An artificial-intelligence approach to computer-assisted instruction. *IEEE Transactions on Human-Machine Systems*, *11*, 190–202.

Carvalho, T., Costa, C., Mombach, J., Ferreira, C., Fernandes, D., & Soares, F. (2020). Petasystem–a piano expression teaching aid system. In *2020 IEEE Canadian conference on electrical and computer engineering (CCECE)* (pp. 1–5). IEEE.

Chaudhari, A., Dalvi, O., Ramade, O., & Ambawade, D. (2021). Yog-guru: Real-time yoga pose correction system using deep learning methods. In *2021 international conference on communication information and computing technology (ICCICT)* (pp. 1–6). IEEE.

Chen, H.-T., He, Y.-Z., Chou, C.-L., Lee, S.-Y., Lin, B.-S. P., & Yu, J.-Y. (2013). Computer-assisted self-training system for sports exercise using kinects. In *2013 IEEE international conference on multimedia and expo workshops (ICMEW)* (pp. 1–4). IEEE.

Chen, H.-T., He, Y.-Z., Hsu, C.-C., Chou, C.-L., Lee, S.-Y., & Lin, B.-S. P. (2014). Yoga posture recognition for self-training. In *International conference on multimedia modeling* (pp. 496–505). Springer.

Chen, H.-T., He, Y.-Z., & Hsu, C.-C. (2018). Computer-assisted yoga training system. *Multimedia Tools and Applications*, *77*, 23969–23991.

Di Mitri, D., Schneider, J., Trebing, K., Sopka, S., Specht, M., & Drachsler, H. (2020). Real-time multimodal feedback with the CPR tutor. In *International conference on artificial intelligence in education* (pp. 141–152). Springer.

Dittakavi, B., Bavikadi, D., Desai, S. V., Chakraborty, S., Reddy, N., Balasubramanian, V. N., Callepalli, B., & Sharma, A. (2022). Pose tutor: An explainable system for pose correction in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3540–3549).

D'Mello, S., & Graesser, A. (2009). Automatic detection of learner's affect from Gross body language. *Applied Artificial Intelligence*, *23*, 123–150.

D'Mello, S., Picard, R. W., & Graesser, A. (2007). Toward an affect-sensitive autotutor. *IEEE Intelligent Systems*, *22*, 53–61.

D'mello, S. K., & Kory, J. (2015). A review and meta-analysis of multimodal affect detection systems. *ACM Computing Surveys (CSUR)*, *47*, 1–36.

Freedman, R., Ali, S. S., & McRoy, S. (2000). Links: What is an intelligent tutoring system? *Intelligence*, *11*, 15–16.

Görer, B., Salah, A. A., & Akın, H. L. (2017). An autonomous robotic exercise tutor for elderly people. *Autonomous Robots*, *41*, 657–678.

Graesser, A. C., Lu, S., Jackson, G. T., Mitchell, H. H., Ventura, M., Olney, A., & Louwerse, M. M. (2004). Autotutor: A tutor with dialogue in natural language. *Behavior Research Methods, Instruments, & Computers*, *36*, 180–192.

Huang, F., Zhou, Y., Yu, Y., Wang, Z., & Du, S. (2011). Piano ar: A markerless augmented reality based piano teaching system. In *2011 third international conference on intelligent human-machine systems and cybernetics, Vol. 2* (pp. 47–52). IEEE.

Johnson, D., Dufour, I., Damian, D., & Tzanetakis, G. (2016). Detecting pianist hand posture mistakes for virtual piano tutoring. In *Proceedings of the international computer music conference* (pp. 166–170).

Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, *6*, 1–54.

Joseph, R., Ayyappan, M., Shetty, T., Gaonkar, G., & Nagpal, A. (2022). Befit—a real-time workout analyzer. In *Sentimental analysis and deep learning* (pp. 303–318). Springer.

Kale, G., Patil, V., & Munot, M. (2021). A novel and intelligent vision-based tutor for yogāsana: E-yogaguru. *Machine Vision and Applications*, *32*, 1–17.

Lara, O. D., & Labrador, M. A. (2012). A survey on human activity recognition using wearable sensors. *IEEE Communications Surveys and Tutorials*, *15*, 1192–1209.

Liu, Q., Yu, S., Wang, Y., Le, H., & Yuan, Y. (2017). A hand-waving dance teaching system based on kinect. In *International conference on blended learning* (pp. 354–365). Springer.

Long, C., Jo, E., & Nam, Y. (2022). Development of a yoga posture coaching system using an interactive display based on transfer learning. *Journal of Supercomputing*, *78*, 5269–5284.

Majumdar, R., Bhawar, P., Sahasrabudhe, S., & Dinesan, P. (2014). Hasta: Hasta training application learning theory based design of Bharatanatyam hand gestures tutor. In *2014 IEEE 14th international conference on advanced learning technologies* (pp. 642–643). IEEE.

Masala, I. V., & Angdresey, A. (2017). The real time training system with kinect: Trainer approach. In *2017 international conference on soft computing, intelligent system and information technology (ICSIIT)* (pp. 233–237). IEEE.

Mat Sanusi, K. A., Mitri, D. D., Limbu, B., & Klemke, R. (2021). Table tennis tutor: Forehand strokes classification based on multimodal data and neural networks. *Sensors*, *21*, 3121.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Group*, P. (2009). Preferred reporting items for systematic reviews and meta-analyses: The prisma statement. *Annals of Internal Medicine*, *151*, 264–269.

Molloy, W., Huang, E., & Wünsche, B. C. (2019). Mixed reality piano tutor: A gamified piano practice environment. In *2019 international conference on electronics, information, and communication (ICEIC)* (pp. 1–7). IEEE.

Mousavinasab, E., Zarifsanaiey, N., R. Niakan Kalhori, S., Rakhshan, M., Keikha, L., & Ghazi Saeedi, M. (2021). Intelligent tutoring systems: A systematic review of characteristics, applications, and evaluation methods. *Interactive Learning Environments*, *29*, 142–163.

Movva, P., Pasupuleti, H., & Sarma, H. (2022). A self learning yoga monitoring system based on pose estimation. In *International conference on human-computer interaction* (pp. 81–91). Springer.

Muangmoon, O-o., Sureephong, P., & Tabia, K. (2017). Dance training tool using kinect-based skeleton tracking and evaluating dancer's performance. In *International conference on industrial, engineering and other applications of applied intelligent systems* (pp. 27–32). Springer.

Munshi, A., Rajendran, R., Ocumpaugh, J., Biswas, G., Baker, R. S., & Paquette, L. (2018). Modeling learners' cognitive and affective states to scaffold SRL in open-ended learning environments. In *Proceedings of the 26th conference on user modeling, adaptation and personalization* (pp. 131–138).

Olney, A. M., D'Mello, S., Person, N., Cade, W., Hays, P., Williams, C., Lehman, B., & Graesser, A. (2012). Guru: A computer tutor that models expert human tutors. In *International conference on intelligent tutoring systems* (pp. 256–261). Springer.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., et al. (2021). The prisma 2020 statement: An updated guideline for reporting systematic reviews. *Systematic Reviews*, *10*, 1–11.

Papastratis, I., Chatzikonstantinou, C., Konstantinidis, D., Dimitropoulos, K., & Daras, P. (2021). Artificial intelligence technologies for sign language. *Sensors*, *21*, 5843.

Patil, S., Pawar, A., Peshave, A., Ansari, A. N., & Navada, A. (2011). Yoga tutor visualization and analysis using surf algorithm. In *2011 IEEE control and system graduate research colloquium* (pp. 43–46). IEEE.

Rajendran, R., Iyer, S., & Murthy, S. (2018). Personalized affective feedback to address students' frustration in its. *IEEE Transactions on Learning Technologies*, *12*, 87–97.

Rajšp, A., & Fister, I. (2020). A systematic literature review of intelligent data analysis methods for smart sport training. *Applied Sciences*, *10*, 3013.

Ramadijanti, N., Fahrul, H. F., & Pangestu, D. M. (2016). Basic dance pose applications using kinect technology. In *2016 international conference on knowledge creation and intelligent computing (KCIC)* (pp. 194–200). IEEE.

Rastgoo, R., Kiani, K., & Escalera, S. (2021). Sign language recognition: A deep survey. *Expert Systems with Applications*, *164*, Article 113794.

Rho, S., Hwang, J.-I., & Kim, J. (2014). Automatic piano tutoring system using consumer-level depth camera. In *2014 IEEE international conference on consumer electronics (ICCE)* (pp. 3–4). IEEE.

Rigby, L., Wünsche, B. C., & Shaw, A. (2020). Piarno-an augmented reality piano tutor. In *32nd Australian conference on human-computer interaction* (pp. 481–491).

Rishan, F., De Silva, B., Alawathugoda, S., Nijabdeen, S., Rupasinghe, L., & Liyanapathirana, C. (2020). Infinity yoga tutor: Yoga posture detection and correction system. In *2020 5th international conference on information technology research (ICITR)* (pp. 1–6). IEEE.

Ritschel, H., Seiderer, A., & André, E. (2020). Pianobot: An adaptive robotic piano tutor. In *Workshop on exploring creative content in social robotics at HRI 2020*.

Ros, R., Coninx, A., Demiris, Y., Patsis, G., Enescu, V., & Sahli, H. (2014). Behavioral accommodation towards a dance robot tutor. In *Proceedings of the 2014 ACM/IEEE international conference on human-robot interaction* (pp. 278–279).

Sharma, P., & Anand, R. S. (2021). A comprehensive evaluation of deep models and optimizers for Indian sign language recognition. *Graphics and Visual Computing*, *5*, Article 200032.

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, *6*, 1–48.

Shrestha, A., & Mahmood, A. (2019). Review of deep learning algorithms and architectures. *IEEE Access*, *7*, 53040–53065.

Sinatra, G.M., Heddy, B.C., & Lombardi, D. (2015). The challenges of defining and measuring student engagement in science.

Sun, C.-H., & Chiang, P.-Y. (2018). Mr. piano: A portable piano tutoring system. In *2018 IEEE XXV international conference on electronics, electrical engineering and computing (INTERCON)* (pp. 1–4). IEEE.

Trejo, E. W., & Yuan, P. (2018). Recognition of yoga poses through an interactive system with kinect device. In *2018 2nd international conference on robotics and automation sciences (ICRAS)* (pp. 1–5). IEEE.

Voulodimos, A., Doulamis, N., Doulamis, A., & Protopapadakis, E. (2018). Deep learning for computer vision: A brief review. *Computational Intelligence and Neuroscience*, *2018*.

Wadhawan, A., & Kumar, P. (2021). Sign language recognition systems: A decade systematic literature review. *Archives of Computational Methods in Engineering*, *28*, 785–813.

Wang, J., Chen, Y., Hao, S., Peng, X., & Hu, L. (2019). Deep learning for sensor-based activity recognition: A survey. *Pattern Recognition Letters*, *119*, 3–11.

Wu, Y., Lin, Q., Yang, M., Liu, J., Tian, J., Kapil, D., & Vanderbloemen, L. (2021). A computer vision-based yoga pose grading approach using contrastive skeleton feature representations. In *Healthcare: Vol. 10* (p. 36). MDPI.

Xu, M., Zhai, Y., Guo, Y., Lv, P., Li, Y., Wang, M., & Zhou, B. (2019). Personalized training through kinect-based games for physical education. *Journal of Visual Communication and Image Representation*, *62*, 394–401.