American Journal of Advanced Multidisciplinary Innovation and Research (AJAMIR)

E-ISSN: XXXX-XXXX ● Website: www.ajamir.com ● Email: editor@ajamir.com

Multimodal Artificial Intelligence for Early Identification of Learning Difficulties

Dr. Thabo M. Nkosi

Senior Lecturer, Southern Institute for Learning Sciences, Cape Town, South Africa

Abstract

Early learning difficulties are frequently recognized only after persistent academic underachievement becomes visible, by which time students may already have experienced repeated frustration, reduced academic confidence, inappropriate instructional placement, or delayed specialist support. Artificial intelligence offers an emerging opportunity to complement teacher observation and conventional educational assessment by detecting patterns across multiple forms of learner data. This study examines the potential of multimodal artificial intelligence for the early identification of learning difficulties through a structured integrative review and conceptual analytical methodology.

Multimodal systems can combine academic responses with eye movements, oral reading characteristics, handwriting behavior, interaction logs, video-derived behavior, physiological signals, and contextual educational information, thereby providing a broader representation of learning processes than single-source screening. Contemporary research demonstrates promising applications in reading difficulties, dyslexia, dysgraphia, engagement analysis, collaborative learning, and adaptive education. A 2025 systematic review of multimodal learning analytics in K–8 education identified only 14 peer-reviewed empirical studies published between 2011 and 2023 and reported continuing weaknesses in explicit data-fusion methodology, ethics, and transparency.

A recent multimodal dyslexia-screening framework integrating eye-gaze, speech, and handwriting information further demonstrated the technical feasibility of multimodal risk identification. Within that framework, the independently evaluated eye-tracking classifier achieved 92.8% accuracy, an F1 score of approximately 0.93, and an area under the receiver-operating-characteristic curve of 0.99; however, these performance measures apply specifically to the eye-tracking module and should not be interpreted as validation of the complete multimodal system. The present paper proposes a Multimodal Learning-Difficulty Identification Pathway integrating low-burden screening, modality-specific feature extraction, explainable data fusion, uncertainty estimation, teacher review, and referral for formal assessment.

The analysis argues that educational AI should identify patterns requiring additional attention rather than autonomously assign diagnostic labels. Responsible implementation requires representative datasets, age-appropriate consent, privacy protection, cultural and linguistic validation, explainability, human oversight, and ongoing evaluation of false-positive and false-negative consequences. Multimodal AI may therefore become a valuable component of inclusive education when it strengthens—not substitutes for—professional educational and clinical judgment.

Keywords: multimodal artificial intelligence; learning difficulties; learning disabilities; early identification; dyslexia screening; learning analytics; eye tracking; handwriting analytics; speech analysis; inclusive education

1. Introduction

Learning difficulties can emerge through diverse patterns involving reading accuracy, decoding, spelling, handwriting, mathematical reasoning, language processing, attention, memory, or the ability to translate knowledge into academic performance. Early identification is educationally important because delayed recognition may postpone appropriate instruction and support. Conventional screening frequently depends on standardized achievement measures, specialist assessment, teacher observation, parent reports, and repeated evidence of academic difficulty. These remain essential because learning difficulties are context-sensitive and cannot be reduced to a single biological or behavioral signal. Nevertheless, many conventional procedures require time, trained personnel, repeated observation, or access to specialists. Contemporary artificial-intelligence research is therefore examining whether digital behavioral signals can provide an additional layer of early risk detection without replacing formal assessment. Recent school-based dyslexia research continues to emphasize the need for early screening and identifies considerable methodological diversity among existing instruments.

The growing field of multimodal learning analytics provides a technical foundation for this approach. Rather than inferring a learner's state from examination scores or clickstream data alone, multimodal systems combine information derived from different channels, potentially including text, audio, video, eye gaze, movement, physiological sensing, digital pen traces, and educational-system logs. Ouhaichi and colleagues describe multimodal learning analytics as an emerging field in which heterogeneous learner information is integrated to generate richer representations of learning processes. Caskurlu, Ocak, and Dai's 2025 K–8 reviews similarly concluded that multimodal analytics has meaningful potential in younger educational populations; although only 14 empirical K–8 studies met their criteria and methodological reporting remained inconsistent. The comparatively limited K–8 evidence base is particularly significant because early-identification technologies are most consequential when used with children.

Reading difficulty provides one of the clearest examples of how multimodality can contribute. Eye tracking can quantify fixation duration, regressions, saccadic behavior, reading speed, dwell time, and scan paths; speech processing can analyze oral-reading speed, pauses, pronunciation errors, and word-error rates; handwriting systems can examine character formation, spacing, reversals, writing speed, pressure, and pen movement. Tiwari and colleagues' 2025 Akshar Mitra framework brought eye tracking, speech recognition, and handwriting analysis within a common digital architecture for early dyslexia screening. Its design processes video-derived gaze, audio-derived reading behavior, and handwriting information as complementary data streams rather than assuming that any individual signal is sufficient. Similar multimodal reasoning can potentially be extended to other forms of learning difficulty, although the empirical evidence is presently stronger for reading- and writing-related difficulties than for many other educational conditions.

The attraction of multimodal AI lies in complementarity. A learner may read slowly because of decoding difficulty, limited vocabulary, anxiety, fatigue, unfamiliarity with the language, or inadequate instruction. Eye-movement information alone may reveal increased processing effort but cannot necessarily identify why it occurs. Speech may reveal decoding or fluency problems, while handwriting can reveal orthographic and graphomotor patterns. Academic response data can indicate whether these difficulties consistently affect learning outcomes, and teacher observation can place the digital signals within the student's classroom history. Multimodal fusion therefore offers an opportunity to search for converging evidence across independent sources. Earlier research in multimodal learning analytics has similarly demonstrated that different learner-data channels can provide complementary information for understanding complex educational behavior and supporting human decision-making.

2. Objectives of the Study

2.1 To Examine Multimodal Indicators of Emerging Learning Difficulty

The first objective is to examine how complementary educational and behavioral signals can contribute to early identification. Relevant modalities include academic-task performance, eye movements during reading, oral-reading speech, handwriting characteristics, digital-learning interactions, observable engagement, and selected physiological or neurocognitive information where ethically and practically appropriate. The purpose is to determine how each modality contributes distinctive evidence while recognizing that no single digital marker should independently establish a learning-disorder diagnosis.

2.2 To Develop an Explainable AI-Supported Screening Pathway

The second objective is to develop a conceptual architecture through which multimodal learner information can be collected, synchronized, analyzed, fused, and converted into interpretable risk indicators. The proposed system emphasizes explainable features, calibrated probabilities, teacher review, longitudinal confirmation, and referral for professional assessment rather than opaque binary labeling.

2.3 To Define Ethical and Educational Requirements for Implementation

The third objective is to identify the requirements necessary for safe adoption in schools. These include informed and age-appropriate consent, data minimization, security, transparency, linguistic and cultural validation, accessibility, teacher training, fairness assessment, uncertainty reporting, and clear restrictions on the educational consequences that may follow an automated risk flag. UNESCO's contemporary AI-in-education framework similarly places human agency and responsible governance at the center of educational AI adoption.

3. Review of Literature

3.1 Multimodal Learning Analytics and Learner-State Identification

Multimodal learning analytics developed from recognition that conventional educational datasets often capture only a restricted representation of learning. Clicks, test scores, and completion times provide useful information but cannot fully represent communication, attention, movement, affect, interaction, or strategy. Mu, Cui, and Huang's systematic review of multimodal data fusion showed that multimodal learning analytics can integrate varied learner information and potentially provide a more accurate understanding of complex learning processes. Ouhaichi, Spikol, and Vogel subsequently mapped the development of the field and identified growing research across learning processes, technologies, modalities, and learning contexts.

Multimodal AI has already been applied to collaborative learning, game-based learning, embodied interaction, and learning-state prediction. Cukurova, Kent, and Luckin demonstrated how AI and multimodal information could support rather than replace human educational decision-making in a debate-tutoring context. Emerson and colleagues showed the potential of multimodal analytics within game-based learning, while Olsen and colleagues demonstrated the value of incorporating temporal multimodal patterns when predicting collaborative learning outcomes. These studies establish a broader methodological precedent: educationally meaningful patterns may emerge from relationships among several information streams rather than from isolated measurements.

3.2 Eye Movement, Speech, and Handwriting as Early-Risk Signals

Reading-related learning difficulties are well suited to multimodal analysis because reading involves coordinated visual, linguistic, phonological, motor, and cognitive processes. Eye-tracking research has repeatedly reported differences in fixation patterns and reading behavior between learners with and without reading difficulties. A 2026 exploratory study involving elementary students found statistically significant differences in average fixation behavior between students with and without identified reading disabilities, reinforcing the potential of eye movement as a screening signal while also emphasizing the need for larger validation studies.

Recent AI research has expanded this approach. Tiwari and colleagues used the publicly available ETDD70 dataset containing eye-tracking recordings from 70 Czech schoolchildren aged 9–10 years, divided evenly between diagnosed dyslexic and typical-reader groups. Their eye-tracking classifier achieved 92.8% accuracy, an F1 score of 0.93, and an AUC of 0.99 on its reported evaluation. Importantly, these results concern the eye-tracking classifier within a larger framework and should not be generalized as the accuracy of the complete multimodal system.

3.3 From Automated Detection to Inclusive Educational Support

Early identification has little educational value if it ends with classification. The purpose of screening should be to initiate timely observation, appropriate instruction, targeted support, or professional assessment. A child flagged because of unusual gaze behavior should not immediately be labeled dyslexic; the system should instead indicate that observed patterns warrant additional investigation.

The Akshar Mitra framework explicitly follows this principle. Its authors state that the application is not a diagnostic tool and position it as an early-screening and support system. The architecture combines screening modules with reading assistance, syllable-level support, behavioral tasks, and gamified learning resources. This distinction is important because machine-learning performance under controlled datasets does not automatically establish clinical or educational validity across real classrooms.

Personalized-learning research similarly suggests that multimodal analytics may support adaptation by providing richer information about individual learning processes. Khor, Tan, and Chan's systematic review specifically examined the application of multimodal learning analytics to personalized learning and argued that multiple learner-data sources may overcome limitations associated with personalization based on a single type of information. The most appropriate future architecture may therefore connect **b**, while reserving formal diagnostic decisions for qualified professionals.

4. Research Methodology

4.1 Research Design

The present study employed a structured integrative evidence-synthesis and conceptual framework-development approach. No original schoolchildren were recruited and no new learner dataset was generated. The methodology was selected because the research problem spans learning analytics, artificial intelligence, inclusive education, special education, reading research, speech analysis, handwriting assessment, eye tracking, and educational ethics.

Peer-reviewed and authoritative evidence available up to August 10, 2026 was considered. Priority was given to empirical educational studies, systematic and scoping reviews, recent peer-reviewed AI screening studies, and UNESCO guidance concerning responsible educational AI.

4.2 Multimodal Data Architecture

The proposed system incorporates low-burden educational information before considering more intrusive sensing methods.

Table 1: Proposed Multimodal Data Architecture for Early Learning-Difficulty Identification

Data Modality	Representative Features	Potential Learning Signal	Primary Limitation
Academic-task data	Accuracy, response time, error pattern	Persistent skill difficulty	Influenced by instruction and prior knowledge
Eye tracking	Fixations, regressions, saccades, dwell time	Reading-processing difficulty	Hardware/calibration and task sensitivity
Speech	Reading rate, WER, pauses, pronunciation	Fluency and phonological difficulty	Accent and language bias
Handwriting	Formation, spacing, pressure, timing, reversals	Written-language or graphomotor difficulty	Device and language dependence
Interaction logs	Attempts, hints, task abandonment, latency	Learning-strategy difficulty	Context dependent
Video/behavior	Gaze direction, posture, observable interaction	Engagement or task difficulty	Privacy and interpretation risk
Teacher observations	Classroom performance and longitudinal context	Ecological confirmation	Observer variation
Screening assessments	Phonological, numerical, memory measures	Domain-specific risk	Requires validated instruments
Physiological data	EEG or other selected signals	Experimental neurocognitive evidence	Cost, intrusiveness, scalability

5. Analysis

5.1 Functional Contribution of Individual Modalities

The available evidence indicates that individual modalities capture different manifestations of learning difficulty. Eye tracking is particularly useful for examining reading behavior; speech can measure fluency, errors, and timing; handwriting can reflect written-language and graphomotor patterns; learning logs can identify repeated difficulties during authentic educational tasks.

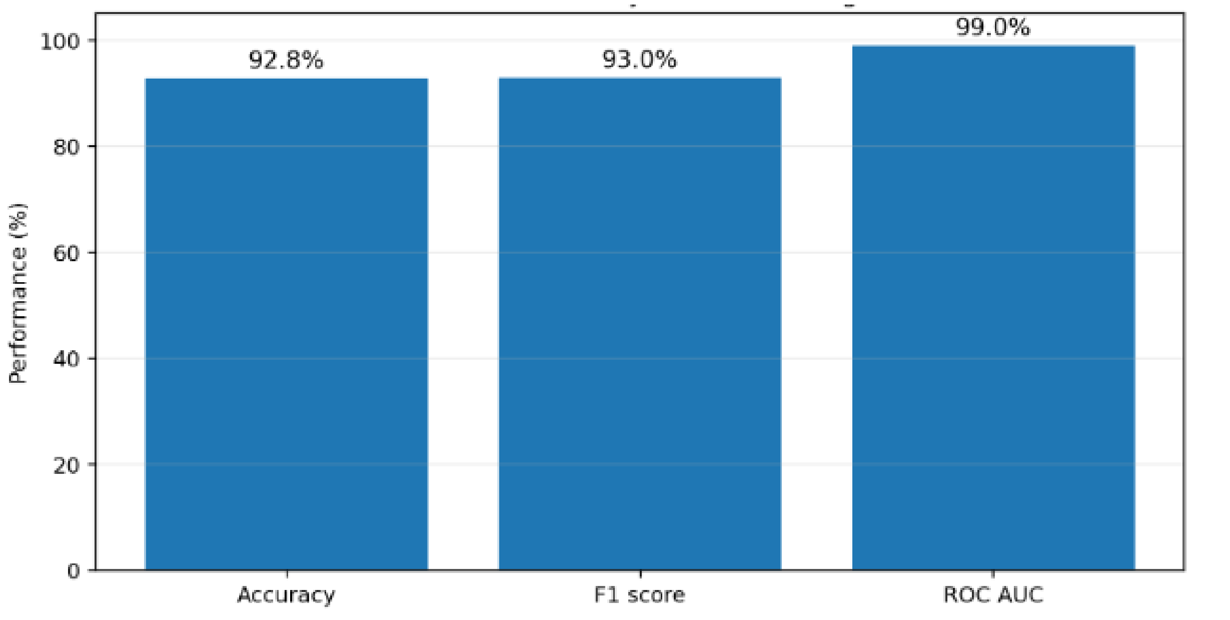
Table 2: Evidence-Informed Interpretation of Multimodal Screening Signals

Observed Pattern	Relevant Modality	Possible Educational Interpretation	Appropriate Response
Longer fixations and repeated regressions	Eye tracking	Increased effort during reading	Additional reading assessment
Slow oral reading with frequent pauses	Speech	Fluency/decoding concern	Phonological and fluency screening
High word or pronunciation error rate	Speech	Possible decoding difficulty	Language-sensitive assessment
Irregular spacing or letter formation	Handwriting	Possible writing/graphomotor difficulty	Handwriting evaluation
Repeated task errors despite practice	Learning logs	Persistent concept difficulty	Targeted instruction
Rapid abandonment and extensive hints	Interaction logs	Difficulty, uncertainty, or low engagement	Teacher review
Multiple convergent abnormal signals	Multimodal fusion	Elevated risk requiring investigation	Formal screening/referral
Single isolated abnormal signal	One modality	Possible noise or contextual effect	Monitor before escalation

5.2 Published Validation Evidence

Tiwari and colleagues provide a contemporary example of quantitative AI screening. Their eye-tracking pipeline was developed using the ETDD70 dataset of 70 schoolchildren, and the classifier reported 92.8% accuracy, an F1 score of 0.93, and and ROC AUC of 0.99.

Figure 1: Reported Performance of the Eye-Tracking Classifier Within a 2025 Multimodal Dyslexia-Screening Framework



5.3 Proposed Multimodal Early-Identification Pathway

The synthesis supports the following operational model:

Routine Classroom Learning Evidence

↓

Teacher or Digital Early-Risk Signal

↓

Domain-Specific Low-Burden Screening

↓

Eye / Speech / Writing / Interaction Features

↓

Explainable Multimodal Fusion

↓

Risk Probability + Uncertainty Estimate

↓

Teacher and Specialist Review

↓

Targeted Instruction / Monitoring / Formal Referral

↓

Longitudinal Reassessment

The model deliberately places human review after algorithmic analysis and before consequential classification. AI functions as an evidence-organizing layer rather than a final authority.

6. Discussion

6.1 Why Multimodal AI May Improve Early Identification

The principal advantage of multimodal AI is that learning difficulties are multidimensional. A reading difficulty may manifest simultaneously through slow reading, unusual eye-movement patterns, repeated decoding errors, reduced comprehension, and specific speech characteristics. A unimodal system sees only one manifestation.

Multimodal systems can potentially reduce dependence on any single noisy measure by examining whether several independent observations tell a coherent story. Tiwari and colleagues explicitly argue that integrating gaze, speech, and handwriting can create a more comprehensive screening profile and reduce reliance on one potentially erroneous signal. Broader MMLA research likewise emphasizes that diverse modalities provide a more holistic representation of learner behavior and can support human educational decision-making.

The approach may also enable earlier intervention because digital signals can potentially be collected during short educational tasks rather than waiting for severe academic failure. However, early identification should mean earlier opportunity for support, not earlier permanent labeling.

6.2 Accuracy Alone Is Not Sufficient

High accuracy is attractive in AI research, but educational screening requires additional metrics. Sensitivity determines how many genuinely at-risk students are identified; specificity concerns how many students without the difficulty avoid unnecessary flagging; positive predictive value depends on the prevalence of the condition in the screened population; calibration determines whether predicted probabilities correspond with observed risk.

False positives can cause anxiety, stigma, unnecessary referrals, or inappropriate expectations. False negatives can delay needed intervention. Consequently, model evaluation must report confusion matrices, subgroup performance, confidence intervals, calibration, and external validation rather than presenting single headline accuracy.

Generalizability is especially challenging across languages. Speech patterns, orthographic structure, reading conventions, handwriting characteristics, and educational expectations vary substantially across cultural contexts. Tiwari and colleagues explicitly identify linguistic diversity and insufficiently representative datasets as obstacles to broader AI-based dyslexia screening.

A model validated with one language, age group, curriculum, or device should therefore not be automatically transferred to another population.

6.3 Privacy, Fairness, and Human Oversight

Multimodal educational assessment can generate highly sensitive learner profiles. Video may reveal identity; voice recordings contain biometric characteristics; gaze data may reveal attention patterns; handwriting is personally identifiable; educational logs provide longitudinal information about academic behavior.

UNESCO emphasizes that educational AI must protect learners' rights, preserve human agency, and incorporate privacy and ethical safeguards. These principles are particularly important for minors, who may have limited ability to understand the long-term consequences of persistent educational data collection.

Schools should therefore apply data minimization. If reading accuracy and speech provide sufficient screening evidence, physiological monitoring should not be introduced merely because it is technically possible. Data should be retained only as long as necessary, access should be controlled, and families should understand which signals are collected and how predictions may influence educational decisions.

7. Conclusion

Multimodal artificial intelligence offers a promising approach to the early identification of learning difficulties because it can integrate complementary information that conventional single-source screening may overlook. Eye movements can provide objective information about reading behavior; speech analysis can reveal fluency and decoding patterns; handwriting analytics can characterize writing processes; digital-learning logs can show persistent task difficulty; and teacher observations can place these signals within authentic educational context.

Recent evidence demonstrates technical promise. The 2025 Akshar Mitra framework integrated eye tracking, speech, and handwriting within a common architecture, and its eye-tracking classifier reported 92.8% accuracy, an F1 score of approximately 0.93, and an AUC of 0.99 within its evaluation setting. At the same time, the 2025 systematic review of K–8 multimodal learning analytics identified only 14 empirical studies and found important deficiencies concerning fusion methodology, ethics, and transparency. The field should therefore be characterized as promising but not yet mature enough for unrestricted automated classification of children.

The most defensible role for multimodal AI is early risk screening and educational decision support. Algorithms should organize evidence, identify unusual patterns, quantify uncertainty, and assist educators in deciding whether additional instruction, monitoring, or specialist assessment is warranted. They should not independently diagnose dyslexia, dysgraphia, dyscalculia, or other learning disorders.

Future research should prioritize large longitudinal datasets, multilingual and cross-cultural validation, external school-based testing, explainable multimodal fusion, fairness analysis, comparisons with validated screening instruments, prospective intervention studies, and evaluation of whether earlier AI-supported identification genuinely improves educational outcomes.

The success of multimodal AI should ultimately be measured not by how many students an algorithm can classify, but by whether it enables educators to recognize genuine learning needs earlier, respond more precisely, and provide support without unnecessary labeling, surveillance, or exclusion.

References

1. S. Caskurlu, C. Ocak, and C.-P. Dai, “The scope of multimodal learning analytics in K–8: A systematic review,” *Journal of Learning Analytics*, vol. 12, no. 2, pp. 224–236, 2025, doi: 10.18608/jla.2025.8505.

2. V. Tiwari, O. Agarwal, M. Sharma, R. Sahu, R. Babar, R. Geddam, M. Awais, and H. Ghayvat, "Akshar Mitra: A multimodal integrated framework for early dyslexia detection," *Frontiers in Digital Health*, vol. 7, 2025, doi: 10.3389/fdgth.2025.1726307.
3. H. Ouhaichi, D. Spikol, and B. Vogel, "Research trends in multimodal learning analytics: A systematic mapping study," *Computers and Education: Artificial Intelligence*, vol. 4, 100136, 2023, doi: 10.1016/j.caeai.2023.100136.
4. S. Mu, M. Cui, and X. Huang, "Multimodal data fusion in learning analytics: A systematic review," *Sensors*, vol. 20, no. 23, 6856, 2020, doi: 10.3390/s20236856.
5. M. Cukurova, C. Kent, and R. Luckin, "Artificial intelligence and multimodal data in the service of human decision-making: A case study in debate tutoring," *British Journal of Educational Technology*, vol. 50, no. 6, pp. 3032–3046, 2019, doi: 10.1111/bjet.12829.

   Emerson, E. B. Cloude, R. Azevedo, and J. Lester, "Multimodal learning analytics for game-based learning," *British Journal of Educational Technology*, vol. 51, no. 5, pp. 1505–1526, 2020, doi: 10.1111/bjet.12992.
6. J. K. Olsen, K. Sharma, N. Rummel, and V. Aleven, "Temporal analysis of multimodal data to predict collaborative learning outcomes," *British Journal of Educational Technology*, vol. 51, no. 5, pp. 1527–1547, 2020, doi: 10.1111/bjet.12982.
7. E. T. Khor, L. P. Tan, and S. H. L. Chan, "Systematic review on the application of multimodal learning analytics to personalize students' learning," *AsTEN Journal of Teacher Education*, Special Issue, 2024.
8. H. Ouhaichi, B. Vogel, and D. Spikol, "Exploring design considerations for multimodal learning analytics systems: An interview study," *Frontiers in Education*, vol. 9, 1356537, 2024, doi: 10.3389/feduc.2024.1356537.
9. M. Berns-Conner et al., "Identifying reading disabilities through eye movements," *Frontiers in Education*, 2026.
10. S. Takahata et al., "Comprehensive search for assessment indicators that reflect handwriting characteristics associated with dysgraphia," *Scientific Reports*, 2025.
11. V. Babushkin et al., "Analyzing handwriting legibility through hand kinematics," *Frontiers in Artificial Intelligence*, 2025.

    R. W. Sait et al., "Lightweight hybrid transformers-based dyslexia detection using multimodal data," *Scientific Reports*, 2025.
12. R. Bakhti et al., "A narrative review of school-based screening tools for dyslexia," *Frontiers in Public Health*, 2025.
13. M. F. Shaikh et al., "Serious gaming and eye-tracking for the screening, monitoring and rehabilitation of specific learning disabilities," *Frontiers in Bioengineering and Biotechnology*, 2025.
14. UNESCO, *Guidance for Generative AI in Education and Research*. Paris, France: UNESCO, 2023.
15. UNESCO, *AI and Education: Protecting the Rights of Learners*. Paris, France: UNESCO, 2025.
16. UNESCO, "Artificial intelligence in education," UNESCO Digital Education Programme, accessed Aug. 10, 2026.