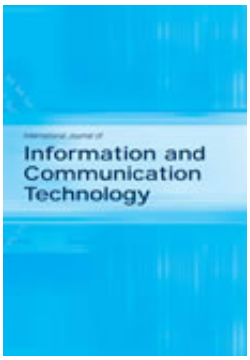


International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Towards an enhanced evaluation framework for English reading competence: leveraging multimodal learning analytics

Lina Liu

DOI: 10.1504/IJICT.2025.10074867

Article History:

Received:	30 June 2025
Last revised:	23 July 2025
Accepted:	23 July 2025
Published online:	17 December 2025

Towards an enhanced evaluation framework for English reading competence: leveraging multimodal learning analytics

Lina Liu

School of Foreign Studies,
Liaoning University of International Business and Economics,
Dalian, 116052, China
Email: llnsbc2025@163.com

Abstract: As the field of educational assessment is growing, traditional ways of testing English reading ability cannot adequately show all the different kinds of information that learners use when they read. Because of this, how to employ multimodal learning behaviour data to make more accurate assessments is a popular topic in educational research right now. This research suggests the MLB-ERAM model for assessing English reading proficiency based on facts on how people learn in different ways. MLB-ERAM uses a lot of multimodal learning behaviour data and deep learning (DL) technology to get a whole picture of how well students can read. The experimental results reveal that the MLB-ERAM model works well with multimodal data, gets around the problems with standard assessment methods, and is a useful guide for the future growth of educational assessment technology.

Keywords: multimodal data; learning behaviour; English reading proficiency assessment; DL.

Reference to this paper should be made as follows: Liu, L. (2025) 'Towards an enhanced evaluation framework for English reading competence: leveraging multimodal learning analytics', *Int. J. Information and Communication Technology*, Vol. 26, No. 46, pp.1–19.

Biographical notes: Lina Liu received her Master's degree from Liaoning University in 2007. She is currently a Lecturer in the Liaoning University of International Business and Economics. Her research interests include English teaching methodology and intercultural communication.

1 Introduction

1.1 Background of study

The area of education is going through an unparalleled phase of digital transformation since information technology is growing so quickly, especially with the widespread use of big data, artificial intelligence (AI), and DL technologies. Standardised examinations, manual scoring, and a single evaluation are often used to measure how well someone can read English. These tests can show how well children understand language to some level, but their results are generally constrained by how subjective the scoring is and how

narrow the evaluation parameters are. Traditional methods of testing do not do a good job of looking at all the different things pupils do while they read, like how they get information, how they read, and how they think about what they read. Also, standardised examinations usually cannot keep track of how pupils are learning in real time, and the feedback is slow and does not include a full study of how well students can read (von Hippel, 2024).

In recent years, with the continuous development of AI technology, especially the progress in the field of multimodal learning behaviour analysis, scholars have begun to try to achieve a more accurate assessment of ability through students' learning behaviour data. Multimodal learning behaviour data includes data generated by students' interactions with the learning content in various ways during the learning process, such as eye movements, keyboard input, mouse clicks, and voice responses (Sharma and Giannakos, 2020). These data can help researchers gain an in-depth understanding of the learning process, thus effectively making up for the shortcomings of traditional research methods.

Nevertheless, how to effectively integrate and analyse multimodal data is still a great challenge. Firstly, students' learning behaviours are highly individualistic, with large differences in the responses of different students to the same reading tasks; secondly, multimodal data tends to be highly dimensional and heterogeneous, making the processing and analysis of the data more complex (Yilmaz et al., 2021). Traditional assessment methods are unable to handle these complex multidimensional data; therefore, new algorithms and models need to be developed to achieve effective data integration and analysis.

This study suggests an English reading competency assessment methodology based on research about how people learn in different ways to tackle the challenges listed above. This paper's goal is to build a more accurate and scientific assessment system by merging different types of perceptual data and employing both DL and data mining methods. Based on this, we not only look at how to effectively combine data, but we also look at how to judge students' English reading skills based on how they learn in real time. This will give us a theoretical framework and practical advice for individualised learning. This paper will go into further detail on how to make the assessment model more accurate and useful, and it will also look at how it might be used in the future in education.

1.2 Significance of study

Standardised tests and manual scoring are often used in traditional ways to test how well someone can read English. These approaches can show how good students are at something to some level, but they typically have issues such being too subjective, not covering enough areas of assessment, and giving feedback too late. These approaches cannot adequately show how children learn and think when they are reading, and they cannot keep track of how students are doing in real time, which makes it hard to fairly judge how well they can read. In the realm of education, we need a more complete, precise, and real-time strategy right away to get over these Humphrey's problems.

The main point of this work is to go beyond the limits of existing evaluation models by adding multimodal learning behaviour data. By combining different types of behavioural data from students while they are learning, like eye movement, keyboard input, mouse clicks, vocal responses, and more, we can get a full and dynamic picture of

how students learn. These data not only show how kids read and absorb material, but they also show how students think and how well they grasp things which gives teachers more personalised feedback on how well their students are learning. The assessment method that uses multimodal data not only gets over the problems with standard assessments, but it can also help with personalised instruction in a big way.

This article also employs data mining and deep learning (DL) to handle and analyse this multimodal data. Data mining can find useful characteristics and patterns in huge amounts of learning behaviour data, and DL can completely explore the possible patterns in the data and make the assessment model better at predicting what will happen. Combining the two makes it possible to get a more realistic picture of how well kids can read English, without the problems that come up with traditional approaches that include people which makes sure that the assessment results are objective and scientific.

This study is also essential since it gives us fresh ideas on how to use educational technology in the future. Combining data mining with DL can help educational decision makers come up with more scientific ways to instruct by giving them solid data backing (Alshadoodee et al., 2022). The evaluation model that uses multimodal data also gives us a theoretical framework for building personalised learning and smart educational environments. This paper is not only very useful for academics, but it also gives useful information for coming up with new ideas and using educational technology in real life.

2 Relevant technologies

2.1 Methods of assessing English reading skills

English reading proficiency tests have come a long way since they first started. They started off as paper-and-pencil examinations, then moved on to computerised tests, and now they use intelligent assessment systems.

Computer-based testing (CBT) has been a popular way to test English reading skills as technology has advanced. CBT, implemented online, allows automatic scoring and quick feedback during testing, improving assessment efficiency. CBT has some drawbacks despite its effectiveness and uniformity. CBT typically uses multiple-choice and fill-in-the-blank questions, which do not effectively test students' complicated text comprehension (Saikh et al., 2022). It still uses pre-set questions, making it hard to reflect pupils' cognitive processes when reading. Computerised evaluation also fails to dynamically capture students' reading techniques, comprehension levels, and affective shifts.

As machine learning (ML), natural language processing (NLP), and DL technologies advance in the 21st century, more study is turning to intelligent assessment approaches. Through large-scale data training, ML helps the evaluation system analyse student behaviour and cognitive features throughout learning. Support vector machines (SVMs) are extensively used to classify kids' reading performance into tiers, while random forest (RF) enhances prediction accuracy by merging several decision trees (DTs) and is good at handling high-dimensional data. By comparing student resemblance to known samples, k-NN algorithms assist computers measuring learning progress and reading techniques. DT algorithms also analyse students' learning behaviours to predict their reading achievement.

DL methods have revolutionised English reading evaluation. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) excel at analysing complicated textual and behavioural data. CNNs are used for text classification and sentiment analysis to correctly measure students' comprehension by extracting representative characteristics from raw text. RNNs and their variations, long and short-term memory networks (LSTMs) and the gated recurrent unit (GRU), simulate pupils' information processing and memory abilities to expose their thought flow and cognitive structure during reading (Yin et al., 2023). These DL models allow the assessment system to examine students' basic understanding, reasoning, and emotive reactions in complex reading assignments.

Even if the current methods used in evaluation systems have come a long way, there are still certain problems that need to be solved. First, most of the ways we assess things now still use data from only one modality. It is still hard to figure out how to combine learning behaviour data from several modalities. To fully show how students think as they read, data fusion needs not only good feature extraction algorithms but also strong data processing and analysis tools. Second, even while DL and ML methods are better for processing massive amounts of data and making accurate predictions, their black-box nature is still a big problem (Indolfi et al., 2025). Models need to be easy to comprehend in the field of educational assessment, and instructors and educators need to know how models make decisions to use assessment results to choose the best ways to teach. So, one major area of current study is how to make evaluation models easier to understand, especially DL models.

Also, more and more people want personalised learning. Current evaluation systems can tell how well pupils are doing overall, but they do not always consider how different each student is. Because each student uses various cognitive methods and reading techniques while they are learning, it is still an essential area of research for English reading assessment to figure out how to give students personalised feedback and assessments based on their unique traits. For instance, transfer learning (TL) technology can assist the assessment system quickly adjust to the needs of each student, make it less reliant on vast sets of annotated data, and make the system more adaptable and flexible.

In the last few years, multimodal learning has slowly become a popular way to test how well people can read in English. By combining information from diverse data sources, multimodal learning can make assessments more complete. Eye-tracking technology, for example, can assist with figuring out how kids pay attention to text and how they read by showing how their attention is distributed and how they absorb information. Speech recognition technology can also help check how well pupils can speak and how well they know the language. This gives more varied data to help with the reading ability test.

In brief, the current methods for assessing English reading competency have fixed some of the difficulties with the old model, but they still need to figure out how to better combine different technologies and make personalised assessments more accurate. This would not only help improve English reading proficiency tests, but it will also strongly support personalised learning and new ideas in education.

2.2 Analysis of multimodal learning behaviour data

Since its emergence, English reading proficiency assessment has gone through several stages of evolution, from the early traditional paper-and-pencil tests to computer-assisted assessment, and up to today's intelligent assessment systems.

Multimodal learning behaviour data analysis struggles with multimodal data fusion. Since modality data have multiple forms, temporal qualities, and sizes, fusing them to derive significant information are problematic in study. Multimodal data fusion has three main types: early, late, and intermediate. Early fusion methods jointly process data from different modalities in the input stage and merge them into a unified feature vector before feeding them into the learning algorithm for training. This method takes advantage of correlation between modalities, but it is also susceptible to noise. However, late fusion approaches process each modality's data individually and merge their outputs. Late fusion has the benefit of allowing each modality to be trained on its own, which means that it is not affected by noise from a single modality (Nagrani et al., 2021). However, the downside is that the independence of multiple modalities could mean that information is lost. Mid-term fusion approaches combine data from several modalities at the intermediate stage of processing, usually after feature extraction. This method keeps the independent features of each modality and improves the fusion process through training, which makes it very flexible. As DL technology has advanced, especially with the use of CNN and RNN, automatic feature extraction and multimodal data fusion have become faster and more accurate. Deep neural network (DNN) can now automatically learn useful feature representations from data of different types through a multilayered network structure.

Despite significant progress in multimodal learning data analysis in many fields, it still faces a number of challenges. Firstly, the heterogeneity of data is a key issue; data from different modalities differ greatly in format, scale, and temporal sequence, etc. and how to integrate them into a unified framework is a technical challenge in multimodal data analysis. Secondly, the data synchronisation problem is also very important, especially when it comes to time series data, how to ensure the time alignment and synchronisation of different modal data to avoid the impact of time errors on the analysis results is the core problem to be solved. Another problem is choosing the right features and reducing the number of dimensions. The dimensionality of multimodal data is usually very high. The key to making the analysis faster and more accurate is to figure out how to get rid of unnecessary redundant information while keeping the information's integrity. The choice of feature extraction and dimension reduction methods will have a direct impact on the effectiveness of data fusion. Finally, model interpretability is also an important issue facing multimodal learning behaviour data analysis (Baltrušaitis et al., 2018). In sensitive areas such as education and healthcare, researchers and applicators need to be able to understand the basis of the model output to ensure transparency and fairness in decision-making.

Despite the above challenges, the future of multimodal data analysis remains promising with the continuous development of technology. The rise of augmented reality (AR) and virtual reality (VR) technologies opens new ways to gather and analyse multimodal data, which makes the behavioural data of learners more detailed and realistic. Cross-modal learning will also become an important area of study in the future which lets a model learn the data features of one modality using the data of another modality. This not only makes multimodal data analysis more useful, but it also makes multimodal fusion faster and more accurate.

In short, multimodal learning behaviour data analytics can change the way we think about education, healthcare, psychology, and human-computer interaction. It can also give us new tools and approaches for better understanding how people think and act.

Even if the subject currently has certain technical problems, multimodal analysis will become more precise and useful as related technologies keep changing and improving. It can be very helpful in many areas.

3 English reading proficiency assessment model

3.1 Collection of experimental data

The dataset utilised in this study is a self-made multimodal learning behaviour dataset that includes text, speech, eye movement, and video data. It was made to give the model in this paper all the information it needs. The dataset was built with the help of 100 college students majoring in English who were between the ages of 18 and 22 and had a particular degree of English reading skills. The university's multimedia classrooms were the major places where data was collected. This was done to make sure that a variety of behavioural data were collected in a real learning environment.

Each participant was required to complete an English reading task and answer relevant comprehension questions after reading in the experiment, during which records in four areas including text data, speech data, eye movement data and video data were collected.

Specifically, the text data consisted of students' multiple-choice and short-answer answers in the reading tasks, which were able to reflect students' comprehension level and mastery of the text. The text content of each task covers a wide range of fields such as science and technology, literature, and history, which ensures the diversity of the data. Information such as students' reading reaction time and answer accuracy was also recorded simultaneously.

Participants will be asked to retell the key points of the article once they finish reading it for the phonological data. Their speech during the retelling will be recorded. These speech statistics can show things like how well children speak the language, how fast they speak, and how their intonation changes. They can also help figure out how mentally and emotionally taxing the reading experience was for them. We used high-quality microphones to record all the speech data so that it would be clear and accurate.

The Tobii Pro eye-tracking gadget collects eye-movement data by recording things like where students are looking, how long they are looking, and how they are scanning the text while they are reading (Wang et al., 2021). These eye-movement data can show how much attention the students paid to different parts of the text while they were reading, and they can also be used to look more closely at the students' reading tactics and cognitive load. Each experimental job takes about 10 minutes to read, which is meant to be like a genuine reading situation. Participants will do the reading task numerous times to get a full picture of how the pupils read.

On the other hand, video data caught students' facial expressions and body movements while they were reading with a high-definition camera. This camera could see how students' emotions changed and how they acted without saying anything. Students' facial expressions, like smiling and frowning, can show how they are feeling, like anxious, happy, or confused. These statistics enable us to look more closely at how students' emotions vary at different stages of reading and how these changes affect their ability to learn.

Table 1 shows the exact makeup of the dataset and how it was gathered.

Table 1 Dataset composition and collection methods

Data type	*Description*	*Collection method*	*Source*
Text data	Responses from students on reading tasks (multiple-choice, short answer)	Online student responses	100 English major undergraduates from a university
Speech data	Speech data generated by students during reading aloud or retelling	Recorded via microphone	Speech recordings from English major undergraduates
Eye-tracking data	Eye movement trajectories, including fixation points and duration	Eye-tracking device (e.g., Tobii Pro)	Collected in a multimedia classroom at a university
Video data	Video recordings of students' facial expressions and body movements during reading	Video camera recordings	Recorded in a multimedia classroom at a university

After gathering all the data, it will be cleansed and pre-processed. Text data will be cleaned up by getting rid of extra information and standardising answer formats. Speech data will be cleaned up by getting rid of background noise and converting it to standard audio formats. Eye movement data will be cleaned up by getting rid of invalid gaze points and correcting trajectories. Video data will be taken from key frames for facial expression changes and sentiment analysis. The data labels will be based on students' behaviour, such as how their emotions change and how much cognitive load they have. These will be paired with the results of the students' reading ability test.

This study's dataset has a lot of different types of multimodal learning behaviour data that span all aspects of students' text comprehension, language proficiency, attention allocation, and emotional changes. The dataset is a good source of information for later model training, competency testing, and personalised feedback. It can also give a lot of information about how learners think, feel, and learn. All data collection was done in a way that respected ethical rules, and the participants gave their consent.

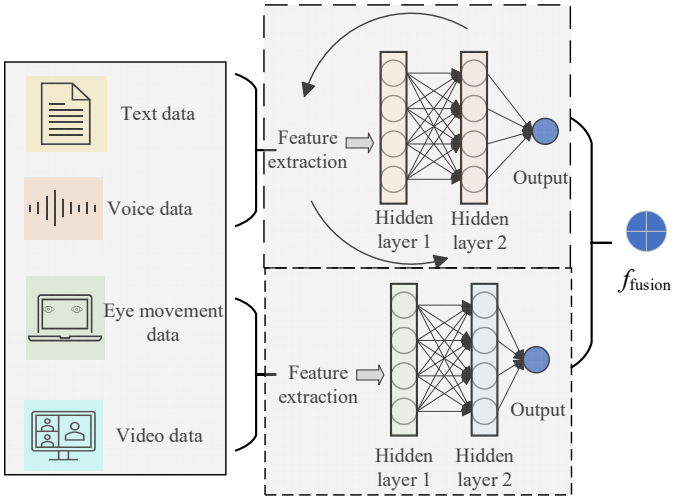
3.2 Design of model

The MLB-ERAM model for assessing English reading proficiency that this paper proposes aims to give students a more accurate and complete picture of their English reading skills by combining data mining and DL techniques with multimodal learning behaviour data, as shown in Figure 1.

3.2.1 Multimodal data acquisition and pre-processing

The main job in the MLB-ERAM model is to gather behavioural data from learners from many different sources and then carefully pre-process this data to make sure that the analysis that comes next is accurate and useful. The model uses multimodal learning behaviour data, which includes text, audio, eye movement, and video data. Each type of data gives a different view of how learners read English.

Figure 1 Design of MLB-ERAM (see online version for colours)



NLP methods are used to process the text data and find the learners' grammatical structures, language properties, and emotional patterns while they read. Word segmentation, stop word removal, lexical annotation, and sentiment analysis are all parts of text pre-processing (Mehanna and Mahmuddin, 2021). These procedures help the behavioural analysis show how well the learners speak, understand what they read, and respond emotionally.

A speech recognition system uses speech data to find important things about the learner's verbal expression, like how fast they speak, how they sound, and how their emotions change. Some of the steps involved in pre-processing speech data are filtering out noise, breaking up speech into segments, extracting audio features, and classifying emotions. These methods help to accurately record how the learner's speech varies during different reading activities, which in turn shows how well they understand what they're reading and how their emotions alter (Song and Park, 2021).

Eye-movement data utilise eye-tracking technology to record the location of the learner's gaze, the path of visual movement and the duration of each gaze during the reading process. When pre-processing this kind of data, it is necessary to smooth the eye-tracking trajectories and remove outliers first. The processed eye-movement data can reveal, to a certain extent, learners' attention distribution, information processing style, and their depth of comprehension of the reading content (Liu, 2018).

Video data, on the other hand, is mainly derived from the capture of learners' facial expressions and body movements, which can reflect their emotional changes and non-verbal behaviours during the reading process. The pre-processing of such data includes steps such as image denoising, extraction of key facial feature points, and model-based emotion classification. These processing tools help to identify the learner's emotional state, level of concentration, and his/her psychological responses when facing the reading material.

When all multimodal data acquisition is complete, they also need to be unified and standardised to ensure that data from different sensors or devices are consistent and comparable in subsequent fusion analysis. Especially for time-series type data, there is

often the problem of inconsistent timestamps due to different acquisition frequencies and start times. The specific implementation method is as follows:

$$\hat{X}_t = f(X_t) \tag{1}$$

$$\hat{Y}_t = g(Y_t) \tag{2}$$

$$\hat{Z}_t = h(Z_t) \tag{3}$$

The data of each modality are X_t, Y_t, Z_t, t denotes the time point, and f, g, h are the interpolation functions of the data of different modalities, which ensure that the data of all modalities can be aligned under the same timestamps and ensure the compatibility and usability of the data.

3.2.2 Feature extraction and data mining

In the MLB-ERAM model, the core objective of the feature extraction and data mining module is to identify representative features from multimodal learning behaviour data and analyse them in depth with the help of data mining techniques, to provide a reliable basis for the subsequent assessment of reading ability. The module integrates multiple analysis methods to extract key information from different modalities, such as text, speech, eye movement and video, and strives to comprehensively portray the cognitive state and behavioural characteristics of learners in the reading process.

First, when processing text data, it mainly relies on NLP technology to conduct in-depth analyses of the learner's linguistic expression characteristics, syntactic structure, and emotional responses. Specific steps include lexical processing, removal of meaningless words, lexical annotation, and emotional tendency judgement, etc. to accurately capture the fluency of language expression, the complexity of syntactic structure, and the emotional characteristics embodied in the process of expression. The extracted text features can be expressed as:

$$f_{text} = \{W, S, C, E\} \tag{4}$$

where W denotes lexical features, S is syntactic structure features, C is sentence complexity, and E is emotional tendency features.

Feature extraction of speech data focuses on audio features such as speech rate, intonation, pauses, etc. which can reveal learners' language fluency and emotional fluctuations. In MLB-ERAM model, Mel frequency cepstrum coefficient (MFCC) technique is used for speech feature extraction to analyse the learner's speech performance by extracting the time-frequency features of the speech signal (Abdul and Al-Talabani, 2022). The formula for MFCC feature extraction is as follows:

$$f_{speech} = \{R, T, P, F\} \tag{5}$$

where R is the speech rate, T is the intonation, P is the pause duration, and F is the emotion fluctuation feature.

The gaze point, gaze length, and scanning path are the main features that are extracted from the eye movement data. These can show how the learners pay attention to different parts of the text and how they absorb information while reading. Denoising and trajectory

smoothing are part of the processing of eye movement data to make sure it accurately shows how learners read (Eskenazi, 2024). We can say the eye movement features as:

$$f_{eye} = \{A, D, V, L\} \tag{6}$$

where A is the glance point, D is the gaze length, V is the speed of the sweep, and L is the long-time stare feature.

Facial expression recognition and motion capture are used in feature extraction of video data to show how the learners' emotions and thoughts change while they read by looking at things like their facial expressions and body language. CNN looks at facial expressions and body languages to figure out how the learner is feeling and how focused they are (Noroozi et al., 2018). The formula for extracting features from video data is:

$$f_{video} = \{E, F, P, C\} \tag{7}$$

where E is the emotional feature, F is the facial expression feature, P is the body posture feature, and C is the feature that shows how emotion has changed.

After completing the feature extraction, the MLB-ERAM model further analyses these features through data mining techniques to discover potential laws and patterns from them. The data mining method uses cluster analysis and classification algorithms to look at the student behaviour data in detail. This helps find groups of learners who have similar ways of thinking and learning. This model sorts students' behaviours using the K-means clustering algorithm and puts students with similar cognitive patterns into groups for personalised evaluation. The data mining process can be represented as follows:

$$C = Cluster\left(f_{text}, f_{speech}, f_{eye}, f_{video}\right) \tag{8}$$

where C represents the classification result of the learner and the function *Cluster*() represents the clustering process, which identifies the cognitive features and learning strategies of the learner by fusing the modal data.

3.2.3 DL modelling and behavioural analysis

Firstly, CNN, as a powerful image processing and feature extraction tool, is used in the MLB-ERAM model to analyse eye movement data and video data. CNN extracts features such as the distribution of gaze points and gaze patterns in eye movement data through the convolutional layer, to reveal the learner's attentional focus and cognitive strategies. During the processing of eye movement data, CNN can identify the learner's attention to different text regions during reading and further speculate the cognitive changes and critical moments in reading. For instance, if a learner stares at certain parts of the content for a long time, it means they are thinking deeply about or understanding that part of the topic. We can describe the convolutional layer of CNN by extracting features from eye movement data like this:

$$f_{CNN} = Conv2D\left(f_{eye}\right) \tag{9}$$

$$f_{CNN} = Conv2D\left(f_{video}\right) \tag{10}$$

CNN can analyse learners' facial expressions and body language in video data to show how their emotions and cognitive states evolve over time. Changes in facial expressions,

such smiling or frowning, might show how learners feel while they read, like happy, anxious, or confused. CNNs can pick up on these emotional changes by processing the video data in a way that is called convolutional processing which provides we with useful information about how well learners understand what they are watching (Hossain and Muhammad, 2019). CNNs can also look at video data in detail to find changes in feelings, focus, and emotional responses. This gives us a better idea of how learners feel and how their moods vary while they read.

Next, RNN is used to process speech data and text data, especially for modelling time-series data, which can effectively capture learners' cognitive patterns and emotional fluctuations. The main role of RNN in the MLB-ERAM model is to analyse learners' speech patterns, language fluency, intonation changes and so on during the reading process, to reveal their cognitive loads and emotional responses. Through the time series modelling capability of RNN, the model can dig deeper into the cognitive changes and emotional fluctuations of learners in different reading stages.

For instance, the cyclic structure of the RNN models speech data such variations in speech rate, intonation, and pauses. This can show how the learner is feeling and how well they understand. The way that speech speeds up or slows down and the way that intonation rises and falls might show how various students react to the material, which in turn shows how they are feeling or how hard they are having trouble understanding it. At the same time, RNN can capture learners' cognitive load and comprehension process in the text and reveal their thinking patterns and emotional fluctuations in reading through its modelling of temporal dependencies (Santhosh et al., 2024). The formula for RNN processing is denoted as:

$$f_{RNN} = RNN\left(f_{speech}\right) \tag{11}$$

$$f_{RNN} = RNN\left(f_{text}\right) \tag{12}$$

Among them, f_{speech} and f_{text} are feature inputs of speech data and text data, respectively, which can capture the learners' temporal dynamic features in speech and language performance after RNN processing. Through the temporal modelling of RNN, the dynamic features of learners in speech and text can be captured to further reveal their cognitive patterns and emotional responses.

3.2.4 Multimodal data fusion and capacity assessment

In the MLB-ERAM model, the core responsibility of the multimodal data fusion and competence assessment module is to efficiently integrate the features extracted from different modalities and based on which to carry out dynamic competence assessment and personalised feedback generation. The module introduces a dynamic assessment mechanism, which enables the system to make flexible and accurate judgments based on students' real-time behavioural performance and emotional changes during the learning process, thus reflecting their cognitive status and competence development trends more realistically.

Among them, the fusion of multimodal data is one of the key aspects of the whole model. The process aims to synthesise the learners' behavioural characteristics reflected by multiple data sources such as eye movement, video, voice and text. These modalities provide complementary information dimensions from the perspectives of attention

distribution, emotional response, linguistic expression and semantic comprehension, respectively. In order to achieve effective information integration, the model employs a fusion strategy based on weight assignment, whereby different modalities are assigned corresponding influence according to their characteristics to highlight their roles in the assessment system. For example, eye movement data is mainly used to portray the degree of attention concentration, video data is used to capture facial emotional changes, speech data reflects the fluency and cognitive load of oral expression, and text data directly reflects language comprehension and logical expression ability. The specific fusion is carried out through the following formula:

$$f_{fusion} = w_{eye} \cdot f_{eye} + w_{video} \cdot f_{video} + w_{speech} \cdot f_{speech} + w_{text} \cdot f_{text} \tag{13}$$

where f_{fusion} is the fused feature vector; w_{eye}, w_{video}, w_{speech} and w_{text} are the weighting coefficients for the eye movement, video, speech, and text data; f_{eye}, f_{video}, f_{speech} and f_{text} are the feature vectors for each type of data. Weighted fusion can create a full picture of the features of each modality, giving full information for ability assessment.

The approach uses a dynamic evaluation method based on multimodal data fusion. This means that the assessment criteria change based on how well the learners are doing and how their emotions are changing in real time. The approach changes the results of learning assessments in real time by keeping an eye on how students' emotions, attention span, and language skills change over time, and by merging students' behaviours at different stages of learning. For instance, if a student stares at something for a long time, takes more breaks, or speaks more slowly in a given passage, the model can guess that the student is having trouble or is under cognitive stress (SMRL and BRW, 2021). It can therefore lower the assessment scores at that stage to represent the student's current cognitive state. We can write the formula for dynamic assessment like this:

$$S_{dynamic} = f_{fusion} \cdot w_{dynamic} \tag{14}$$

where $S_{dynamic}$ denotes the dynamic assessment score, $w_{dynamic}$ is the weighting coefficients adjusted according to students' behavioural changes and emotional fluctuations, and f_{fusion} is the fused multimodal features.

The integration of multimodal data enhances the comprehensiveness of the assessment, while the combination of dynamic assessment and feedback mechanism enables the system to provide more adaptive support based on learners' actual performance and individual characteristics, thus promoting their continuous progress and personalised development.

4 Experimental design and realisation

4.1 Experimental setup

The first thing to think about is that the learning rate has a big effect on how quickly the model converges. The trials started with a learning rate of 0.001 and utilised the Adam optimiser to improve it (Xue et al., 2022). The Adam optimiser can make training more efficient and stable by automatically changing the learning rate. The size of the batch also determines how often each training update happens. A batch size of 32 was specified, which is a good mix between speed and memory use (Piao et al., 2023). The model went

through 50 training rounds to make sure it could completely learn and improve from the data.

The dropout strategy was used to stop the model from overfitting. It did this by setting a dropout rate of 0.5, which meant that 50% of the neurons would be randomly lost in each training cycle. This made the model better at generalising.

Table 2 shows how the experiment was setup.

Table 2 Specific setup of the experiment

Parameter	*Value*	*Description*
Learning rate	0.001	Initial learning rate for Adam optimiser
Batch size	32	Number of samples in each training batch
Epochs	50	Number of training epochs
Dropout rate	0.5	Dropout rate to prevent overfitting

With the arrangement above, the model may be trained and tested efficiently using multimodal data. We changed the choice of all hyperparameters via several rounds of trials to make sure the results were stable and correct.

4.2 Proportional analysis of the contribution of different modes to model performance

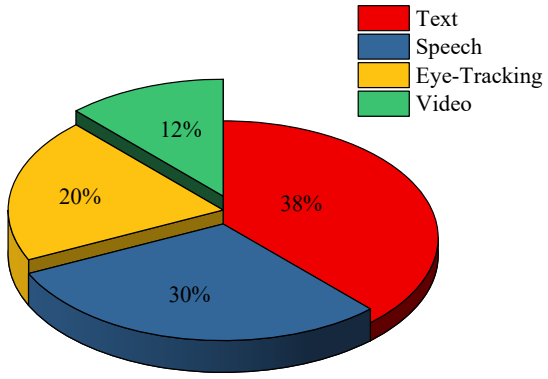
The goal of this experiment is to see how much each of the four types of input (text, audio, eye movement, and video) helps the model work better. The focus will be on how important each type of input is for the model evaluation. Weighted accuracy is the assessment statistic used in this experiment since it better shows how each modality affects the total performance.

The MLB-ERAM model was used in the experiment to train text, speech, eye movement, and video data separately, and the impact of each modality on model performance was calculated. For text data, the model evaluates English readers' comprehension of linguistic content, particularly text semantics and structure. Analysis of phonological data includes the learner's language fluency and phonological qualities including speed, intonation, and pauses, which represent cognitive state and affective reactions. Eye-movement data tracks learners' gaze points to demonstrate their attention distribution and information processing techniques throughout reading. Video data examines learners' emotional fluctuations and cognitive load through facial expressions and body postures, helping to comprehend their psychological condition when reading.

To find out how much each modality added to the weighted accuracy index, we looked at each modality on its own. Figure 2 shows the outcomes of the experiment.

The experimental results show that the contribution of each modality of the MLB-ERAM model to the weighted accuracy is very different. This shows how each modality is used to evaluate the model. The in-depth investigation of the contribution of each modality can further demonstrate the influence of different data sources on model evaluation.

Figure 2 Performance contribution results for different modes (see online version for colours)



First, text data makes up the largest part, at 38%. This shows that text data is the most important part of model assessment. Text data shows how well students understand language, especially when it comes to reading comprehension and extracting information. Text data gives the model the most information about the learners' understanding of syntax, semantics, and other language aspects, as well as their mastery of the content of the text. So, text data is the most significant part of figuring out how well someone can read English.

Phonological data has the next highest contribution, accounting for 30 per cent. Phonological data mainly reflect learners' performance in oral expression in terms of fluency, rate of speech, and intonation, and these features can provide information about the cognitive load of learners in the reading process. Phonological data are crucial for assessing learners' verbal expression, especially in terms of fluency and emotional expression. Therefore, speech data, although failing to surpass the contribution of text data, still occupies a more important position in the model.

Among the various data modalities, the contribution of speech data is relatively high, accounting for 30%. Speech information mainly reflects the learners' characteristics of fluency, speed control and intonation change in oral expression, which can reflect their cognitive load status in the reading process to a certain extent.

The contribution of eye movement data is 20%, which is the third among all modalities. It can reflect the area of concentration of learners' attention while reading and their information processing style, and the cognitive strategies they adopt can be inferred from the eye movement trajectories. Although access to high-quality eye movement data usually requires the support of more specialised equipment, leading to some limitations in its application, it still provides valuable additional information for understanding learners' internal cognitive processes.

Video data made the smallest contribution, accounting for only 12%. In this experiment, video modality has a relatively limited impact on model performance improvement, but it still has some reference value in some specific dimensions, such as emotion recognition and mental state judgement. Although it is not as significant as text and speech data in the current task, video data may become a key source of information in other contexts where visual cues are predominant, such as studies that need to capture facial micro-expressions or body movements.

Taken together, this result provides important insights into multimodal learning behaviour data-driven assessment of English reading ability, suggesting that a

combination of different modalities can lead to more comprehensive and accurate assessment results.

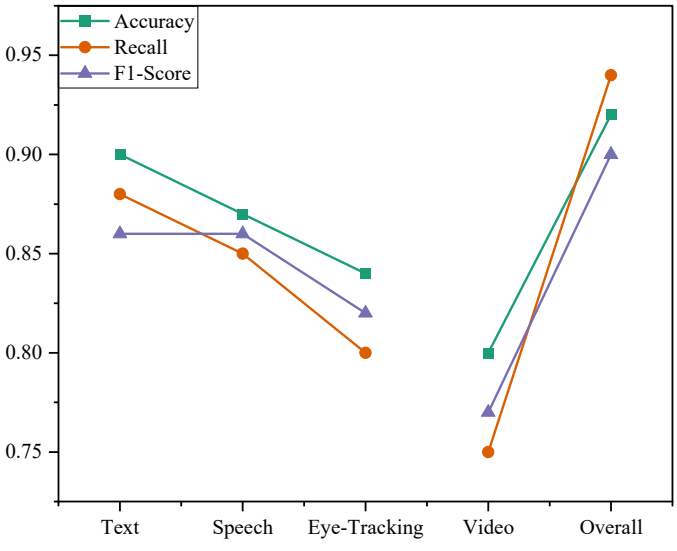
4.3 Performance evaluation with different modal data

A complete performance evaluation experiment was created to evaluate the MLB-ERAM model's performance on various data types. The experiment will examine the model's performance on a multimodal learning behaviour data-driven English reading competence assessment problem, focusing on performance measure variation.

In terms of performance evaluation, metrics such as accuracy, recall, F1-score, and weighted accuracy were used. Through these indicators, we can better understand the performance of the model under different modal data.

The MLB-ERAM model is trained using data from each modality, and the accuracy, recall, and F1-scores are recorded in each modality. The weighted accuracy and final composite scores are calculated by combining the performance of each modality to comprehensively assess its performance. Finally, statistical analyses are used to compare the results across modalities to validate multimodal data fusion and improve its performance. The experimental results are shown in Figure 3.

Figure 3 Performance evaluation of different modes (see online version for colours)



From the experimental results, the performance of the MLB-ERAM model in the four data modalities is significantly different, but the overall performance of the model is significantly improved by fusing these multimodal data. The textual data did the best, with an accuracy of 0.90, a recall of 0.88, and an F1-score of 0.86. This shows how important written data is for judging how well someone can read English, since it may correctly show how well they understand what they read. Text data is the most direct linguistic information on how well learners understand and learn, hence it is the most important type of data.

Voice data did rather well too, with an accuracy of 0.87, a recall of 0.85, and an F1-score of 0.86. Speech data mostly shows how well learners can speak, how fast they can speak, and how well they can use intonation, among other things. Speech data is not very good at directly measuring understanding, but it can be a useful addition to learners' emotional ups and downs and expressive skills. Speech data is vital for figuring out how learners are feeling and how their minds are working, which is why it is such a big part of the whole evaluation system.

The performance of eye-movement data is relatively more general, with an accuracy of 0.84, a recall of 0.80, and an F1-score of 0.82. Eye-movement data can effectively reveal the distribution of the learner's attention and his/her cognitive strategies during the reading process, for example, the learner's gaze time and change of the gaze point in different regions of the text, which reflect the learner's degree of concentration and his/her way of processing the information. However, the complexity of acquiring and processing eye-movement data makes its improvement in overall performance limited. Nonetheless, eye-movement data still provides a necessary addition to the model, especially when analysing learners' concentration and information processing.

Video data contributes little, with an accuracy of 0.80, recall of 0.75, and F1-score of 0.77. By analysing the learner's facial expressions, bodily movements, and other non-verbal behaviours, video data might help the model understand affective swings. Video footage is useful in sentiment analysis; however, the subjective character of sentiment swings and the variety of non-verbal behaviours limit its assessment capabilities in this experiment. Video data can still support learners' emotive states and cognitive load.

Overall, the MLB-ERAM model works really well. The model got an accuracy of 0.92, a recall of 0.94, and an F1-score of 0.90 by combining text, audio, eye movement, and video data. These results show that even while the model does better with one type of data than another, combining different types of data makes it much better at doing whole assessments. Textual data gave the most accurate picture of understanding, while speech, eye movement, and video data added to the overall picture of learners from many angles, such as their changing feelings, cognitive processes, and mental states. MLB-ERAM can fully evaluate learners' skills, get over the problems with single-modal evaluation, and boost overall performance through this multimodal fusion.

In summary, the multimodal data fusion strategy of the MLB-ERAM model significantly improves the accuracy and comprehensiveness of the assessment of English reading proficiency, especially after combining text, speech, eye movement and video data, the model shows a more robust and accurate ability in assessing learners' reading ability.

5 Conclusions

5.1 Summary of study

This study proposes the MLB-ERAM model, which uses text, speech, eye movement, and video data to evaluate English readers. The experimental results reveal that the model outperforms standard methods in all modalities, especially text data processing, and reading comprehension assessment.

The MLB-ERAM model's multimodal data fusion allows it to gather richer learning behaviour information and complement each other across modalities. Text data, as the core information source, provides the model with the most direct basis for evaluation, while speech, eye movement and video data complement the analysis of learners' affective state, attention and language fluency from multiple perspectives. Taken together, the overall performance of the model demonstrates its significant superiority in assessing learners' English reading ability, adapting to the individual needs of different learners and providing accurate ability assessment.

5.2 Limitations and future work

Although the MLB-ERAM model has demonstrated good performance in multimodal learning behaviour data-driven assessment of English reading proficiency, it still has some limitations.

Firstly, the data collection and pre-processing stages still face challenges. The collection and processing of eye-movement data, speech data, and video data are cumbersome and time-consuming, and the accuracy and quality of these data are often limited by the accuracy of the equipment as well as the environment. Despite standardised processing methods, data quality may still be insufficient in some real-world application scenarios, affecting the accuracy of the final assessment. Second, although the fusion of multimodal data improves the performance of models, how to effectively handle and fuse the heterogeneity between different modalities remains an important issue. There are differences in the characteristics and information presentation of different data modalities, and how to perform data fusion without losing key information is a direction that needs to be further optimised in future research (Gao et al., 2020).

Future research can be improved in the following ways. One is to explore more efficient data collection methods to improve the real-time and accuracy of data. For instance, better sensor technology and more accurate algorithms can be utilised to improve the quality of eye movement and speech data. Secondly, in terms of model optimisation, more advanced techniques such as self-supervised learning and RL can be introduced to further improve the adaptability and robustness of the model under different tasks and scenarios (Ericsson et al., 2022). Third, for the fusion of multimodal data, more advanced data fusion algorithms are attempted, which may help to better capture the correlation between different modalities and improve the performance of the model.

Overall, the MLB-ERAM model, which is a multimodal learning behaviour data-driven tool for testing English reading ability, did rather well in this experiment. But if technology becomes better and data processing methods get better, future research should be able to get around the current problems and make the model more accurate, useful, and scalable. This will make educational assessments more accurate and tailored to each student, and it will help smart education and personalised learning grow.

Declarations

All authors declare that they have no conflicts of interest.

References

Abdul, Z.K. and Al-Talabani, A.K. (2022) 'Mel frequency cepstral coefficient and its applications: a review', *IEEE Access*, Vol. 10, pp.122136–122158.

Alshadoodee, H.A.A., Mansoor, M.S.G., Kuba, H.K. and Gheni, H.M. (2022) 'The role of artificial intelligence in enhancing administrative decision support systems by depend on knowledge management', *Bulletin of Electrical Engineering and Informatics*, Vol. 11, No. 6, pp.3577–3589.

Baltrušaitis, T., Ahuja, C. and Morency, L-P. (2018) 'Multimodal machine learning: a survey and taxonomy', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 41, No. 2, pp.423–443.

Ericsson, L., Gouk, H., Loy, C.C. and Hospedales, T.M. (2022) 'Self-supervised representation learning: introduction, advances, and challenges', *IEEE Signal Processing Magazine*, Vol. 39, No. 3, pp.42–62.

Eskenazi, M.A. (2024) 'Best practices for cleaning eye movement data in reading research', *Behavior Research Methods*, Vol. 56, No. 3, pp.2083–2093.

Gao, J., Li, P., Chen, Z. and Zhang, J. (2020) 'A survey on deep learning for multimodal data fusion', *Neural Computation*, Vol. 32, No. 5, pp.829–864.

Hossain, M.S. and Muhammad, G. (2019) 'Emotion recognition using deep learning approach from audio–visual emotional big data', *Information Fusion*, Vol. 49, pp.69–78.

Indolfi, C., Agostoni, P., Barillà, F., Barison, A., Benenati, S., Bilo, G., Boriani, G., Brunetti, N.D., Calabrò, P. and Carugo, S. (2025) 'Expert consensus document on artificial intelligence of the Italian Society of Cardiology', *Journal of Cardiovascular Medicine*, Vol. 26, No. 5, pp.200–215.

Liu, H-C. (2018) 'Investigating the impact of cognitive style on multimedia learners' understanding and visual search patterns: an eye-tracking approach', *Journal of Educational Computing Research*, Vol. 55, No. 8, pp.1053–1068.

Mehanna, Y.S. and Mahmuddin, M. (2021) 'The effect of pre-processing techniques on the accuracy of sentiment analysis using bag-of-concepts text representation', *SN Computer Science*, Vol. 2, No. 4, p.237.

Nagrani, A., Yang, S., Arnab, A., Jansen, A., Schmid, C. and Sun, C. (2021) 'Attention bottlenecks for multimodal fusion', *Advances in Neural Information Processing Systems*, Vol. 34, pp.14200–14213.

Noroozi, F., Corneanu, C.A., Kamińska, D., Sapiński, T., Escalera, S. and Anbarjafari, G. (2018) 'Survey on emotional body gesture recognition', *IEEE Transactions on Affective Computing*, Vol. 12, No. 2, pp.505–523.

Piao, X., Synn, D., Park, J. and Kim, J-K. (2023) 'Enabling large batch size training for DNN models beyond the memory limit while maintaining performance', *IEEE Access*, Vol. 11, pp.102981–102990.

Saikh, T., Ghosal, T., Mittal, A., Ekbal, A. and Bhattacharyya, P. (2022) 'ScienceQA: a novel resource for question answering on scholarly articles', *International Journal on Digital Libraries*, Vol. 23, No. 3, pp.289–301.

Santhosh, J., Pai, A.P. and Ishimaru, S. (2024) 'Toward an interactive reading experience: deep learning insights and visual narratives of engagement and emotion', *IEEE Access*, Vol. 12, pp.6001–6016.

Sharma, K. and Giannakos, M. (2020) 'Multimodal data capabilities for learning: what can multimodal data tell us about learning?', *British Journal of Educational Technology*, Vol. 51, No. 5, pp.1450–1484.

SMRL and BRW (2021) 'Exploring the English learning strategies of an indigenous Papuan student of Indonesia', *The Qualitative Report*, Vol. 26, No. 9, pp.01–2768.

Song, J. and Park, M-H. (2021) 'Emotional scaffolding and teacher identity: two mainstream teachers' mobilizing emotions of security and excitement for young English learners', *International Multilingual Research Journal*, Vol. 15, No. 3, pp.253–266.

von Hippel, P.T. (2024) 'Two-sigma tutoring: separating science fiction from science fact', *Education Next*, Vol. 24, No. 2, p.1.

Wang, Y., Lu, S. and Harter, D. (2021) 'Multi-sensor eye-tracking systems and tools for capturing student attention and understanding engagement in learning: a review', *IEEE Sensors Journal*, Vol. 21, No. 20, pp.22402–22413.

Xue, Y., Tong, Y. and Neri, F. (2022) 'An ensemble of differential evolution and Adam for training feed-forward neural networks', *Information Sciences*, Vol. 608, pp.453–471.

Yilmaz, Y., Aktukmak, M. and Hero, A.O. (2021) 'Multimodal data fusion in high-dimensional heterogeneous datasets via generative models', *IEEE Transactions on Signal Processing*, Vol. 69, pp.5175–5188.

Yin, C., Tang, D., Zhang, F., Tang, Q., Feng, Y. and He, Z. (2023) 'Students learning performance prediction based on feature extraction algorithm and attention-based bidirectional gated recurrent unit network', *PLoS ONE*, Vol. 18, No. 10, p.e0286156.