RESEARCH JOURNAL OF ENGLISH (RJOE)
www.rjoe.org.in | **Oray's Publications** | ISSN: **2456-2696**
An International Approved Peer-Reviewed and Refereed English Journal
Impact Factor: *8.373 (SJIF)* | **Vol. 11, Issue 2 (April/May&June;2026)**

Explainable Deep Transformer Framework for Personalized English Learning Through Multimodal Cognitive and Behavioural Signals

Habab Osman Hassan Ahmed
Language Instructor
Jazan University
English Department
https://orcid.org/0000-0001-9312-1140
E-mail:hohassan@jazanu.edu.sa

Paper Received on 09-03-2026, Accepted on 10-04-2026
Published on 12-04-26; DOI:10.36993/RJOE.2025.11.01.138

Abstract: Deep learning-based personalized systems of learning English tend to use inputs of a single modality, like text or speech, which makes it hard to support the multidimensionality of learner cognition and engagement. Traditional methods often do not take into account physiological cognitive indicators and interaction behavioural dynamics making them less adaptable and less interpretable. In order to overcome these shortcomings, this paper suggests an explainable deep transformer framework to personalize English learning using Multimodal Cognitive and Behavioural Signals. The framework combines three complementary modalities, specifically textual-behavioural data, the Adaptive English Learning Interaction Dataset, which is modelled with the help of RoBERTa, cognitive signals data, the Cognitive Load EEG Dataset, and speech data, the L2-ARCTIC corpus, which are encoded with the help of HuBERT. The modalities are converted to contextual embedding independently and then adapted by a Gated Multimodal Unit (GMU) to fuse modalities. The unified representation is sent to a classification head to predict the learner's performance, and Integrated Gradients gives the interpretability of features by modalities. The model was coded in Python in libraries based on Transformers and trained with cross-entropy loss. Experimental analysis shows good results, with an accuracy of 93.5, F1-score of 0.92, a macro precision of 0.91, and AUC of 0.92, which is better than unimodal baselines. The suggested framework can help the intelligent tutoring system, the educator, and the learners because it facilitates a clear and data-driven personalization accommodating cognitive load, pronunciation fluency, and engagement behaviour, improving the effectiveness of adaptive learning and pedagogical trust.

Keywords: Behaviour Signals; Deep Transformer; English Learning; Multimodal Fusion; Temporal Fusion Transformer

I. INTRODUCTION

The English language learning is becoming increasingly transformed by artificial intelligence with the assistance of chatbots, adaptive learning applications, and natural language processing systems that provide instant feedback and customized instructions. Despite the use of AI in the acquisition of speaking, listening, and writing skills, various challenges have been identified, including cultural limitations, privacy, disparities, and overreliance on technology [1]. Deep learning in language learning involves interrelated competencies like learner motivation, engagement, strategic learning procedures, and directional capabilities. The dimensions in the context of the online English as a Foreign Language (EFL) context should be understood so as to evaluate significant learning processes. Validated models of measurement also focus on the application of credible measurement models in estimating and promoting higher-order cognitive development in the context of digital language learning [2]. The use of Artificial Intelligence in English language learning is transforming it because it enables low-achieving students across different levels to study in a personalized and performance-oriented manner. Intelligent feedback and practice are offered by AI using Natural Language Processing (NLP), machine learning, speech recognition and intelligent tutoring systems. However, the ethical concerns, data security, inequality in access, and preparedness of teachers are also significant issues [3].

The approach to teaching English speaking offers a theoretical framework of creating a successful teaching plan based on visual, auditory, and contextual tools according to the multimodal discourse analysis. Multimodal teaching improves communicative reality, sensory communication, and motivation of learners in the classroom at the elementary level. Higher-order evaluation tools, such as fuzzy measures and speech recognition models, help with effective evaluation of the performance at the phoneme level of speaking [4]. The ability to speak English requires proper pronunciation and effective assessment tools. Conventional speech extraction algorithms are usually complicated and have poor representation ability, which decreases scoring. The multimodal fusion methods involving audio and visual information, more sophisticated sequence modelling, and matching algorithms provide better speech error detection and more realistic pronunciation evaluation [5]. E learning platform has become an important educational resource, which helps in the interactive and adaptive learning settings. This platform supplements the primary education by making use of multimedia content, quizzes, and virtual tools that supplement classroom teaching. It allows individualized learning strategies, follows the academic progress and emotional conditions, and promotes multilingual and multimodal communication. The performance can be tracked by teachers and parents, and engagement, motivation, and individualized learning outcomes can be enhanced with the help of adaptive quizzes and avatar-based interfaces [6].

In order to overcome these shortcomings, the proposed study is an Explainable Deep Transformer Framework of Personalized English Learning based on Multimodal Cognitive and Behavioral Signals. The model combines a RoBERTa-based text encoder to model the writing proficiency and semantic coherence, a HuBERT-based speech encoder to model the pronunciation and fluency representation, and a Temporal Fusion Transformer to model the sequential engagement patterns of behaviour. Such modality-specific embeddings are dynamically combined with a GMU, and they are capable of dynamically weighting the capital of cognitive and behavioural inputs. The Integrated Gradients is used to increase the level of

interpretability by measuring the contribution of features in a single mode or across multiple modalities. The envisioned architecture will facilitate fine-grained representation of the learners, personalization under adaptive conditions and transparent decision support in intelligent English learning systems.

A. Research Motivation

The explainable artificial intelligence methods contribute to transparency and cognitive interpretability, attributing the predictions to particular audio and textual features [7]. Traditional image-based classification models are usually sensitive to noise in the background, and also do not offer much interpretability, which can be used in practical communication support applications in the real world [8]. The growing susceptibility of high-performing text classifiers to fine-scale adversarial examples highlights the need for transparent systems that reveal model sensitivities and decision boundaries. It is possible to use attribution-based explainability techniques to systematically identify influential tokens, which makes it possible to detect adversarial attacks without changing model architectures or training processes [9]. A combination of deep learning methods in evaluation systems would allow for stronger modelling of nonlinear relationships between the variables of education [10]. The feature extraction through deep learning offers a powerful system of modelling the nonlinear relationships and converting heterogeneous inputs into unified representations.

B. Research Significance

The suggested architecture brings intelligent English learning to the next level, filling the gap between cognitive performance modelling and behavioural analytics in a single Transformer-based platform. The system, with the combination of text, speech, and time elements of engagement, helps to record multidimensional representations of learners that are more representative of the educational interactions in the real world. With an integrated Gated Multimodal Unit, adaptive modality weighting becomes possible, which allows a fine-grained personalization of a variety of different learner profiles. Moreover, the application of Integrated Gradients provides better model transparency, enabling to provide comprehensible feedback and reliable decisions. The work will help to develop the next generation of explainable educational AI systems that will balance predictive performance and accountability to students through educational practice, thus contributing to scalable, data-driven, and ethically responsible personalized learning locations.

C. Problem Statement

The current music mood recognition models have difficulty correctly representing emotional content because of the shortcomings of traditional deep learning models and inadequate feature capture of multi-dimensional audio signals. There is a need for a more powerful framework of transformer-based learning that can be developed, which will be able to learn high-level acoustic representations and enhance the performance of mood classification in a variety of categories [11]. The existing speech recognition technology applied in English language learning is struggling with the ability to extract and segment meaningful features of spoken information, resulting in increased transcription errors and decreased customization. The traditional models can be easily incapable of modelling temporal dependencies. To ensure that a learner can be provided with effective feedback, a deep learning architecture that is a hybrid should be developed to improve recognition accuracy and reduce word error rate [12].

The currently available English language learner assessment schemes are either based on black box deep models that are not very interpretable or rule-based schemes with low predictive ability. A severe lack exists in the unification of explainable rule mining with sophisticated transformer designs to create dependable and understandable proficiency quantification of structured and unstructured student information [13]. To overcome these limitations, a transformer-based model is implemented, all the encoders are fused using GMU, and further, predictions are explained using IG.

D. Research Contributions

The proposed model includes the following contributions:

- Presents a multimodal model (Transformer-based) for personalized learning of English by means of combining text, speech, and behavioural cues.
- Presents modality-specific Transformer encoders, such as RoBERTa to textual proficiency modelling, SCTT to EEG cognitive signal processing and HuBERT to speech pronunciation and fluency representation.
- Uses a Gated Multimodal Unit that allows adapting the weights of modality and dynamic fusion of features.
- The use of Integrated Gradients is included to offer interpretable cross-modal attribution analysis.

E. Rest of the section

Section II includes the related work of various authors based on transformer, deep learning and multimodal fusion and their works. Section III introduces the proposed explainable deep transformer framework for personalized English learning using multimodal fusion. Section IV presents the results and discussions, followed by the conclusion and future scope in Section V.

II. RELATED WORKS

Sun [11] suggested an analysis that works on the creation of a feature-centric deep learning framework of music mood recognition, classifying the emotional substance of music by means of transformer-based and mainstream neural designs. The overall concept of the author is to use artificial intelligence to detect moods to maintain applications like recommendations and emotional intelligence systems. This strategy is based on transformer-based AI HuBERT architecture, ConvFormer, LSTM, and the pretrained YAMNet to be evaluated comparatively. All the models operate in order to train patterns between extracted features and emotion labels, and HuBERT incorporates the use of contextualized audio representations to enhance the performance of classification. Findings indicate that HuBERT has the best accuracy, which is better than other methods. However, the datasets are small, features are hand-crafted, which can be subject to bias and the evaluation is carried out based on a mere split, which can reduce generalization.

Orosoo et al. [12] developed a more modern speech recognition-based framework of personalized English language learning, which combines both a multilayer perceptron and a long short-term memory network in enhancing transcription accuracy. The main concept is to improve the online English learning websites: to properly translate the spoken language into

the written one and overcome the constraints of multimodal feature extraction and segmentation that were outlined by previous NLP studies. The approach integrates both MLP to extract discriminative acoustic features and LSTM to model temporal patterns of speech sequence, thus making it possible to identify continuous spoken language. The role of this hybrid MLP LSTM architecture is to represent nonlinear features as well as sequential patterns; thus, it minimizes transcription errors and enhances learning feedback. The system is implemented in Python, and its Word Error is 0.075 with an overall accuracy rate of 98.25%, which is higher than commercial speech-based learning applications. Disadvantages however, include little discussion on dataset scale, may over- fit since very high accuracy, and may not be tested over a wide range of different accents and in noisy real-world conditions.

Zhao [13] proposed a hybrid deep learning and fuzzy logic model of feature-based evaluation of English language learners, which merges interpretability and high predictive accuracy. The key concept is to combine the rule-based reasoning with the fuzzy logic with high-end transformer-based representation learning to develop a complete proficiency assessment framework. It implements a DBML framework of DeBERTa as the backbone transformer to extract contextual textual embeddings, metadata features reflecting demographic and cognitive attributes and LSTM layers to provide a temporal model and integrate dense features. The role of this hybrid structure is to process unstructured textual responses along with structured behaviour data jointly, to project proficiency in language and retain interpretability by fuzzy rule mining. Findings demonstrate higher performance with the highest accuracy of 93% over traditional machine learning and standalone transformer baselines, and are statistically validated and explainable AI methods, including SHAP and DeepSHAP. Nonetheless, its possible disadvantages are complexity of architecture, computing cost and reliance on high-quality metadata to maintain consistent performance with mixed populations of learners.

Fan [14] suggested integrating STEAM philosophy and deep learning to develop personalized high school English course materials that stimulate interdisciplinary learning and innovation. The essence is to use Transformer based framework to automatically produce and deliver instructional content in English and in accordance with the themes of science, technology, engineering, arts, and mathematics. The process is based on an encoder-decoder Transformer architecture, which has been trained on a pre-processed dataset of 300 high school students and is optimized using the Adam optimizer. The role of the model is to acquire semantic and contextual representations of STEAM-integrated English material and produce accurate, versatile, and pedagogically significant teaching resources with varying goals. Experimental findings showed better results with an overall accuracy of 0.87 and higher results in grammatical accuracy, vocabulary relevance, and cultural sensitivity than the baseline models. The disadvantages, however, are that it relies on computational resources and is not scalable in resource-limited learning settings.

Chen et al. [15] introduced a multimodal deep learning system to learn legal English, which combines visual, text, and speech as auxiliary information to improve intelligent education systems, to develop a cross modal legal English question answering system that would be able to perform detailed reasoning tasks that involves image, text, and acoustic sensor data in an integrated vision language speech encoding system through dynamic attention modelling. It uses multimodal fusion and attention-based representation learning, which merge

heterogeneous features and enhance semantic knowledge in legal settings. The experimental findings indicate high accuracy or performance with an accuracy of 0.87, a precision of 0.88, a recall of 0.85, and high satisfaction of the learners. But the disadvantages can be the complexity of the system, the reliance on the quality of sensors, and the possible difficulty of scaling such systems to various legal education settings.

Zeng [16] proposed a behavioural pattern of business English learners analyzed through deep learning to predict their performance and optimize personal learning. The point is to simulate not only the characteristic features of stillness but also the time behaviour of studying to reveal the significant learning patterns and enhance the educational performance. The approach combines Convolutional Neural Networks to learn spatial or fixed behavioural features with Recurrent Neural Networks to learn sequential and time-varying learning processes to create a hybrid architecture. The purpose of such a CNN RNN model is not only to find out correlations among study habits and academic performance but also to allow changing the learning paths dynamically and providing real-time suggestions. The findings indicate that deep learning is an efficient tool to forecast learner performance and increase interactivity with intelligent systems of education. But there are disadvantages, such as the lack of information on the scale of the datasets, the risk of overfitting the behavioural models, and the inability to generalize the results to the whole learner population and institutional setting.

TABLE I. RELATED WORKS OF AUTHORS AND THEIR LIMITATIONS

Author	Methods Used	Advantages	Limitations
Sun [11]	HuBERT, ConvFormer, LSTM, YAMNet, STFT, MFCC	High accuracy 95%, strong audio modelling	Small dataset, simple split, limited generalization
Orosoo et al. [12]	MLP for features, LSTM for sequence	Very low WER 0.075, high accuracy 98.25%	Accent diversity unclear, dataset size not detailed
Zhao [13]	DeBERTa, Metadata, LSTM, Fuzzy logic, SHAP	High accuracy 93%, explainable results	Model complexity, heavy computation

Fan [14]	Transformer encoder-decoder, Adam optimizer	Good content generation, strong accuracy 0.87	High resource need, limited scalability
Chen et al. [15]	Vision language speech encoder, dynamic attention	Strong multimodal QA, high user satisfaction	Complex system, sensor dependency
Zeng [16]	CNN plus RNN hybrid	Captures behaviour patterns, supports personalization	Limited dataset detail, possible overfitting

TABLE I gives a comparative overview of available literature in the proposed study. It brings out major contributions of different authors, the methodologies used, the key benefits realized and the limitations realized.

III. PROPOSED FRAMEWORK FOR PERSONALIZED ENGLISH LEARNING THROUGH MULTIMODAL COGNITIVE AND BEHAVIOURAL SIGNALS

The proposed framework works with the multimodal inputs of the learners, including EEG-based cognitive features, speech records and behavioural interaction logs, that are independently coded by specific Transformer-based encoders. EEG cognitive representations are represented in a multi-head self-attention Transformer to describe inter-feature relationships between spectral and biometric signals. Speech inputs are coded with HuBERT that makes use of Transformer layers to extract contextualized pronunciation and fluency representations. Interaction sequences through behavioural representations are learned through a Transformer encoder to gain temporal engagement patterns. The resulting modality-specific embeddings are then projected to a common latent space and combined dynamically by a GMU, which dynamically learns to apply weights of importance to cognitive and behavioural signals. The fused data is sent to a prediction head, where it gets learner proficiency of engagement classified. Finally, Integrated Gradients is used to calculate feature-level attributions, which can be transparently interpreted as the impact of multimodal attributions on model decisions.

Fig. 1 shows a multi-stage processing pipeline, with three independent streams of inputs: EEG-based cognitive signals, speech recordings of the L2-ARCTIC corpus, and textual or interaction-based data of the Adaptive English Learning dataset. The modalities are then initially processed in the specific way prescribed by the nature of their signal; EEG modalities are standardized, speech modalities have their waveforms normalized, and text modalities are tokenized. The modality-specific Transformer-based encoders are followed by the pre-processed data, RoBERTa for extracting contextual semantic representations of text, the SCTT

models to inter-feature relationship in EEG cognitive signals and HuBERT to generate contextualized acoustic embeddings of speech.

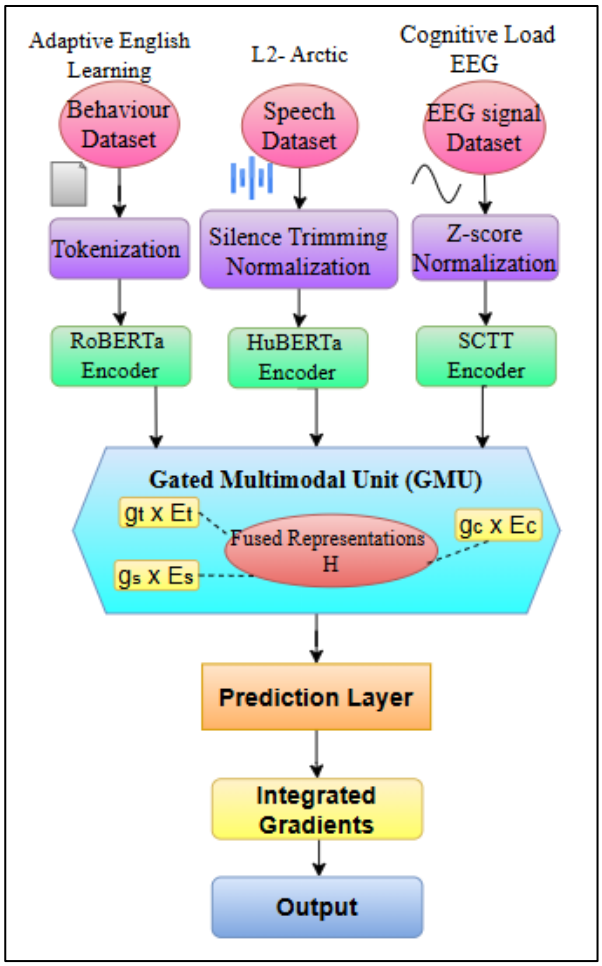

Fig. 1. Block Diagram of the Proposed Model

A. Data Collection and Data Preprocessing.

Three publicly available datasets were used to facilitate multimodal modelling of cognitive and behavioural indicators of individualized English learning and reflect physiological cognition (EEG), linguistic speech proficiency, and learning interactions. The datasets were chosen in a manner that supports diversity of the types of signals and still makes them compatible with Transformer-based encoding architectures. To follow the principles of reproducibility and transparency, data collection was conducted through publicly available research repositories. They are standardized and pre-processed using modal-specific preprocessing that was based on structural characteristics, such as normalization, noise reduction, sequence structuring, and feature scaling. These preprocessing steps provide stability of the numbers, minimize bias due to variance, and normalize the representation of features so that they can later be combined in a multimodal way.

1) Adaptive English Learning Interaction Dataset

The Adaptive English Learning Interaction Dataset was obtained from the publicly available Kaggle dataset, containing records of learners' engagement and performance within a personalized English learning environment [17]. The dataset includes demographic characteristics of the learners, the engagement variables of the learner per session and the assessment scores of the skills like vocabulary, grammar, and reading and the overall performance label.

Before the modelling, an ordinal encoding was used to encode categorical variables like proficiency level and session identifiers in order to make them numerical. Continuous

variables like frequency of engagement, response time and assessment scores have been normalized by the use of min-max scaling, as in (1).

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (1)$$

Where, x is the original feature value, $x_{\min}$and $x_{\max}$denote the minimum and maximum values of the feature within the dataset, and $x'$is the normalized output. This transformation ensures stable training of the behavioural Transformer encoder.

2) Cognitive Load EEG Dataset

The article used to retrieve the Cognitive Load EEG Dataset of English Reading was found on an academic data repository, which is open-access [18]. It includes features of spectral power derived by EEG in Delta, Theta, Alpha, Beta, and Gamma frequency bands, mental effort scores, signal entropy measurements, heart rate variability, and pupil dilation. The sample population was composed of 124 college students who completed reading comprehension tests using English and annotated each reading segment based on the levels of cognitive load (Low, Medium, High).

EEG spectral features and biometric indicators were standardized to remove inter-feature scale disparities using z-score normalization:

$$z = \frac{x - \mu}{\sigma} \quad (2)$$

where $x$is the original feature value, μ is the dataset mean, $\sigma$is the standard deviation, and z is the standardized output.

3) L2-ARCTIC Speech Dataset

L2-ARCTIC corpus was acquired based on an open research speech database of speech records made by non-native English speakers and annotated phonetic alignment, from Kaggle datasets [19]. The sample has speakers of various linguistic backgrounds reading passages in controlled English that allows for examining the accuracy of pronunciation and the variability of fluency. Its organized phonetic correspondence is beneficial in elaborate acoustic modelling in the study of language learning.

The entire recording of the speech was resampled back to 16kHz at 16 bits to have a consistent sampling of the speech. Normalization of amplitude was done to minimize the variation between recording sessions. Through amplitude-based thresholding, silence trimming was done, as in (3).

$$x_t = \begin{cases} 0, & |\, x_t \,| < \tau \\ x_t, & |\, x_t \,| \geq \tau \end{cases} \quad (3)$$

Where, $x_t$denotes waveform amplitude at time t, and τrepresents the silence threshold. The cleaned waveform signals were directly input into the HuBERT encoder for contextual acoustic representation extraction.

B. Proposed Multimodal Transformer Architecture

In this section, the modality-specific Transformer encoders of behavioural-textual signals, physiological cognitive signals and speech-based linguistic signals are described.

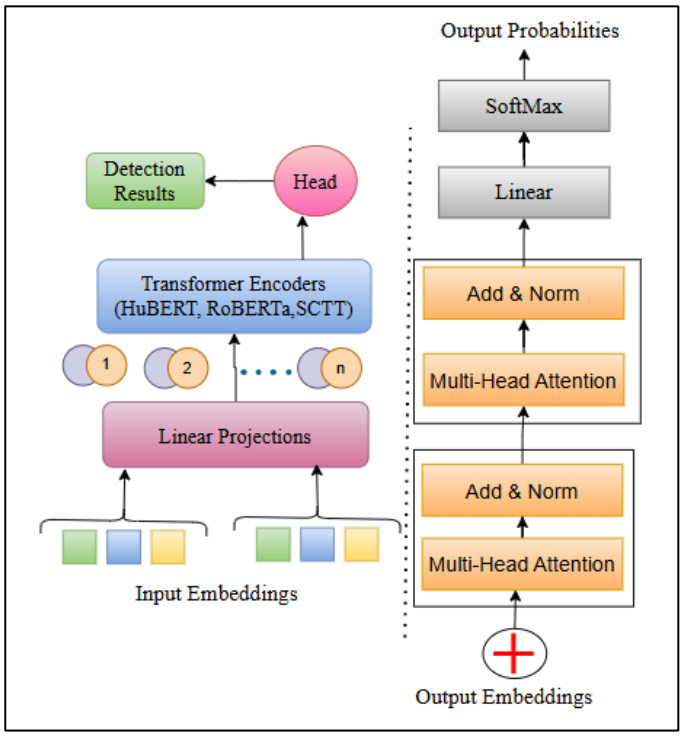

Fig. 2. Proposed Transformer Architecture.

Each encoder transforms raw inputs into contextualized embeddings, which are later fused for prediction. The proposed transformer architecture is given in Fig. 2.

Fig. 2 explains that transformer architecture is a deep learning model which is founded on multi-head self-attention models that allow modelling long-range dependencies in sequential data efficiently. Transformers take inputs in parallel, unlike recurrent models, which enables faster training and learning of better contextual representations. Transformer encoders are used in the proposed framework in all modalities to learn the semantic relationships in text (RoBERTa), inter-feature relations in EEG cognitive signals (self-attention Transformer), and contextual acoustic patterns in speech (HuBERT). Through attention-based feature interaction, the transformer architecture provides consistent and high-capacity representation learning across the diverse cognitive and behavioural inputs, which are the main pillars of a multimodal learning system.

1) Adaptive English Learning (Textual–Behavioural Signals) using RoBERTa

RoBERTa, a highly optimized Transformer encoder architecture that is founded on multi-head self-attention, is used to model the textual constituents of the Adaptive English Learning dataset. RoBERTa eliminates the next-sentence prediction and uses dynamic masking during pretraining, resulting in better representation of contextual language [20].

The input sequence is given in (4) and token embeddings, as in (5).

$$X = \{x_1, x_2, \dots, x_n\} \tag{4}$$

$$E_i = E_i^{token} + E_i^{position} \tag{5}$$

Where, x_i is the i[th] token E_i^{token} is the token embedding and $E_i^{position}$ is the positional encoding. RoBERTa uses multi-head self-attention, as in (6) and (7). The contextualized output embedding is given as in (8).

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (6)$$

$$Q = XW_Q, K = XW_K, V = XW_V \quad (7)$$

$$E_t = \text{RoBERTa}(X) \quad (8)$$

These embedding captures semantic coherence, syntactic structure, and language proficiency patterns.

2) Cognitive Load EEG Dataset using SCTT for Time-Series)

In the application of EEG-based cognitive modelling, the SCTT is used. This architecture directly runs multi-head self-attention on structured multivariate EEG feature vectors, allowing the modelling of multi-spectral band and biometric inter-feature dependencies.

The SCTT uses EEG features, which are based on the conversion of pre-processed cognitive signals to structured spectral representations. Once bandpass filtered and segmented, wavelet or spectral decomposition is used to extract band power features (theta, alpha, beta and gamma) of individual channels at individual time windows [21]. The model subsequently uses two types of self-attention: a temporal attention that characterizes changing patterns of cognitive load across time, and channel attention that characterizes the relationships between electrodes and the dependence between space.

Each EEG segment is represented as:

$$C = \{c_1, c_2, \dots, c_m\}, c_i \in \mathbb{R} \quad (9)$$

$$Z = CW \quad (10)$$

Where, c_i is an EEG feature, and m is the number of EEG features, C is the input EEG feature matrix and $W$is the learnable projection matrix. Multi-head attention is computed as in equations (11),(12) and (13).

$$\text{MHA}(X) = \text{Concat}(head_1, \dots, head_h)W_O \quad (11)$$

$$head_i = \text{Attention}(Q_i, K_i, V_i) \quad (12)$$

$$E_c = \text{SCTT}(C) \quad (13)$$

Where, E_c is the cognitive embeddings. This representation captures dominant EEG frequency interactions and cognitive load indicators.

3) HuBERT-based L2-ARCTIC Speech Modelling

Encoder Overview The HuBERT (Hidden-Unit BERT), a Transformer-based self-supervised speech representation model, is used to process speech signals of the L2-ARCTIC corpus. HuBERT is a convolutional feature extractor [22] in combination with Transformer encoder layers, given in equations (14) and (15).

$$Z = \text{CNN}(X_{audio}) \quad (14)$$

$$E_s = \text{Transformer}(Z) \quad (15)$$

Where, E_s is the speech embedding, Z represents the projected feature representation, and X_{audio} is the raw speech waveform.

C. Multimodal Fusion mechanism.

Once modality-specific embeddings are obtained with the help of RoBERTa (textual behavioural signals), the EEG Transformer encoder (physiological cognitive signals), and HuBERT (speech-based linguistic signals), the following step will be to combine these heterogeneous representations into a single learner profile. As the encoder produces a high-dimensional contextual embedding capturing a different dimension of the learning performance, there is a need to have a fusion mechanism through which the encoders can be combined in a meaningful and adaptive way. The gating vector is computed as in (16).

$$g = \sigma(W_g[\tilde{E}_t; \tilde{E}_c; \tilde{E}_s] + b_g) \quad (16)$$

Where, $W_g \in \mathbb{R}^{d\times 3d}$, $b_g \in \mathbb{R}^d$, and $\sigma(\cdot)$is the sigmoid activation function. The final fused representation is obtained through adaptive weighting as in (17).

$$H = g_t \odot \tilde{E}_t + g_c \odot \tilde{E}_c + g_s \odot \tilde{E}_s \quad (17)$$

Where,g_t, g_c, g_srepresent modality-specific gate components and $\odot$ denotes element-wise multiplication. This mechanism enables dynamic prioritization of cognitive or behavioural signals depending on the learner's state.

A GMU is used to accomplish this, as it is intended to get to know the extent of the importance to assign to each modality as it changes with training. These weights are limited between 0 and 1 with the help of a sigmoid activation, which allows the model to focus on the prominent signals and downplay less informative ones. The last fused representation is calculated as a weighted average of the projected modality embeddings [23]. This adaptive weighting approach enables the model to customize decisions based on the cognitive load the learner is experiencing, the quality of the pronunciation and the behaviour that the learner is using now, thus facilitating a strong multimodal integration.

D. Optimization Strategy and Prediction Layer.

The combined multimodal representation is sent to a fully connected classification head that will forecast the category of performance of the learner. The classification head comprises a sequence of dense layers, then a SoftMax activation function, which generates a probability distribution over the specified output classes. The most likely label is the one that is predicted. The cross-entropy loss function is used to model optimization, and this will assess the difference between predicted probabilities and actual class labels. All model parameters, such as the ones of the modality-specific encoders, fusion module, and classification head, are jointly optimized during training using backpropagation. It uses the Adam optimizer because it has adaptive learning rate characteristics and is stable when training deep Transformer architectures. Mini-batch training is used to enhance the stability of the convergence and computational efficiency. Regularization methods like dropout and early stopping are also implemented to address the issue of overfitting since the architecture is multimodal. The fused representation $H \in \mathbb{R}^d$is passed through a fully connected classification head defined as in (18).

$$\hat{y} = \text{Softmax}(W_o H + b_o) \quad (18)$$

Where, $W_o \in \mathbb{R}^{k \times d}$, $b_o \in \mathbb{R}^k$, and $k$denotes the number of target classes (e.g., proficiency levels). The predicted output $\hat{y}$represents the probability distribution over class labels. Model parameters are optimized using the cross-entropy loss function as in (19).

$$\mathcal{L} = -\sum_{i=1}^{k} y_i \log(\hat{y}_i) \quad (19)$$

Where, y_i is the ground-truth label and $\hat{y}_i$ is the predicted probability for the class i. Training is conducted using mini-batch gradient descent with the Adam optimizer, and backpropagation updates all encoder, fusion, and classification parameters jointly to minimize $\mathcal{L}$.

E. Explainable Artificial Intelligence Mechanism.

Since the application context in education is educational, interpretability is required to provide pedagogical transparency and credibility. Integrated Gradients (IG) is used as an attribution method to give explanations of model predictions. Integrated Gradients calculates the contribution of each input feature by calculating the changes in the prediction of the model when the input changes through a baseline reference point to the real data point. IG uses simpler gradient accumulation, unlike simple gradient methods, producing more reliable and stable attribution scores. In the case of textual inputs, IG identifies influential tokens that have the greatest contribution to the predicted learner state.

In the case of EEG characteristics, it determines predominant spectral bands or biometric indicators of changes in cognitive load. In the case of speech signals, attribution maps are used to show which acoustic frames or segments of pronunciation were used to make the final decision. The interpretability framework allows instructors and learners to know why a specific level of proficiency or engagement was anticipated by offering modality-specific importance scores. This makes the model transparent, facilitates actionable feedback, and makes the use of AI-driven personalized learning systems more ethical. Integrated Gradients (IG) is applied to quantify feature-level contributions across modalities. For an input $x$and a baseline input x', the attribution for the i-th feature is computed as in (20).

$$IG_i(x) = (x_i - x_i') \int_0^1 \frac{\partial F(x' + \alpha(x - x'))}{\partial x_i} d\alpha \quad (20)$$

Where, $F(\cdot)$ denotes the prediction function, x_i is the input feature, and $\alpha \in [0,1]$is the interpolation coefficient.

Algorithm 1: Explainable Deep Transformer for Personalized English Learning Algorithm

Initialize RoBERTa encoder parameters θ_t
Initialize EEG Transformer (SCTT) parameters θ_c
Initialize HuBERT encoder parameters θ_s
Initialize GMU fusion parameters θ_f
Initialize classification head parameters θ_o
Set learning rate η
Set the number of epochs E
Load datasets:
Adaptive English Learning dataset

EEG Cognitive Load dataset
L2-ARCTIC speech dataset
Compute
Normalize behavioural features
Standardize EEG features
Resample and clean speech waveforms
Tokenize textual inputs using RoBERTa tokenizer
While (epoch ≤ E) do
For each mini-batch B in the training data, do
Compute textual embedding
$E_t = \text{RoBERTa}(X_{text})$
Compute EEG cognitive embedding:
$E_c = \text{SCTT}(X_{EEG})$
Compute speech embedding:
$E_s = \text{HuBERT}(X_{audio})$
Compute gating weights:
$g = \sigma(W_g[E_t; E_c; E_s] + b_g)$
Fuse embeddings:
$H = g_t E_t + g_c E_c + g_s E_s$
Compute output probabilities:
$\hat{y} = \text{Softmax}(W_o H + b_o)$
Compute cross-entropy loss:
$\mathcal{L} = -\sum y \log(\hat{y})$
Compute gradients:
$\nabla_\theta \mathcal{L}$
Update parameters:
$\theta = \theta - \eta \nabla_\theta \mathcal{L}$
End For
End While
For each test sample, do
Compute prediction $\hat{y}$
Compute Integrated Gradients attribution:

$$IG_i(x) = (x_i - x_i') \int_0^1 \frac{\partial F(x' + \alpha(x - x'))}{\partial x_i} d\alpha$$

End

Algorithm 1 describes the proposed framework to allow multimodal learning and explainability to be effective. Use of RoBERTa, which is a Transformer-based language model, is used to produce contextualized textual and behavioural inputs as embeddings with the help of multi-head self-attention to extract the semantic dependencies. In the case of cognitive modelling, an encoder is an SCTT that works with both spectral and biometric features derived from EEG data and learns cross-feature interaction through attention processes that have been optimized on structured time-series data.

The SoftMax-based classification head is used to obtain the final prediction, which is optimized with the cross-entropy loss and trained with the Adam optimization algorithm. Also, IG is implemented as an explainability algorithm to compute feature-level attribution scores, which result in transparent and comprehensible model decisions in the cognitive, linguistic, and behavioural domains.

IV. RESULTS AND DISCUSSION

The proposed Explainable Deep Transformer Framework was tested to determine its potential to predict multimodal cognitive and behavioural signals to individually tailor learning English. The model combines RoBERTa to represent text, a Self-Contained Transformer to model EEG-based cognitive, and HuBERT to encode speech-based language, a GMU fusion and explainability of the results by Integrated Gradients. Experimental assessment indicates that multimodal integration is highly rated to predict performance as opposed to unimodal baselines. The adaptive fusion process leads to a successful trade-off between cognitive, behavioural, and speech signal contributions, and superior classification strength and cross-fold generalization.

TABLE II. SIMULATION PARAMETERS

Parameters	Value
Datasets	Adaptive English Learning Interaction Dataset, Cognitive Load EEG Dataset, L2-ARCTIC
OS	Windows 11
Tools	Python version 3.10, PyTorch
Optimizer	Adam
Learning Rate	1e-4
Batch Size	16
Number of Epochs	50
Hidden Dimension (d)	768
Number of Attention Heads	8
Dropout Rate	0.3
Loss Function	Cross-Entropy
Activation Function	SoftMax
Fusion Mechanism	GMU
Explainability	Integrated Gradients

TABLE II shows the simulation parameters, adjusted in a way to achieve stable optimization and efficient convergence of the proposed multimodal Transformer framework. The Adam optimizer was used with a learning rate of 1×10^{-4} to train the model, and it updates the gradient adaptively to deep Transformer models. The batch size of 16 was chosen because it is reasonable to balance the computational efficiency and stable gradients, and to train 50 epochs with early stopping to avoid overfitting. Multi-class classification was done using cross-entropy loss, and SoftMax activation was used at the last prediction layer. Experiments were all conducted using PyTorch, utilizing a graphics card in order to achieve fast training and reproducibility.

A. Data Preprocessing Results

Preprocessing was found to enhance the consistency and stability of convergence and performance of models. Minimizing variance-related gradient instability was achieved by behavioural features that were normalized based on minmax. EEG was standardized to provide equal contribution of spectral bands and biometric characteristics. Z-score normalization brought EEG spectral features to zero with unit variance, scale imbalance was mitigated, and Transformer-based attention modelling was more stable, as in Fig. 3.

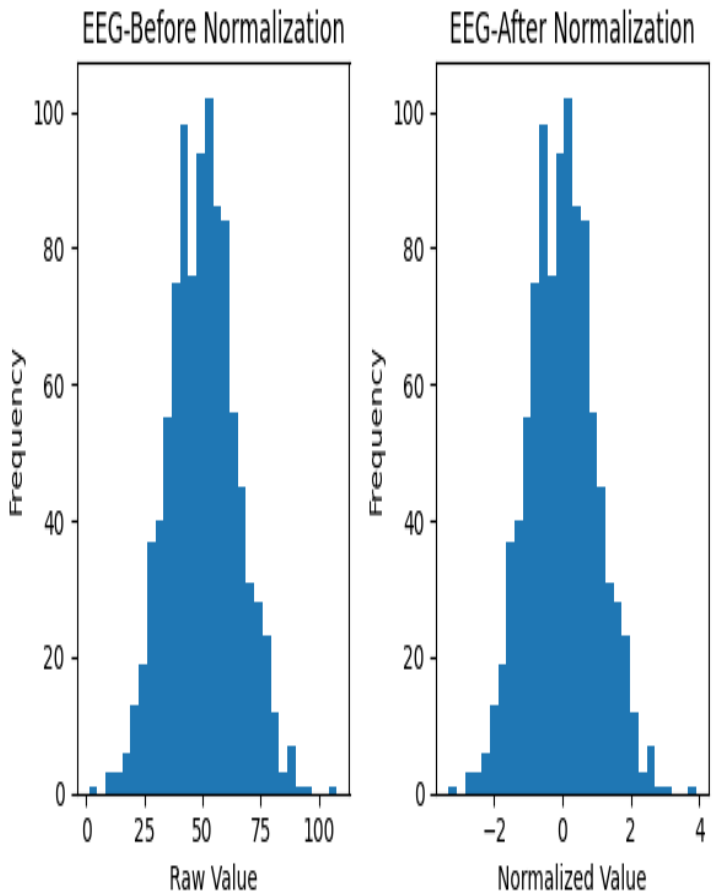

Fig. 3. EEG Feature Before and After Normalization

Fig. 3 shows the distribution of features after preprocessing had lower skew and better inter-feature correlation. The correlation matrix analysis revealed that normalized EEG beta and gamma bands had a moderate correlation with the label of cognitive loads, and engagement measures had a high correlation with end performance classes. Correct preprocessing was also used to help achieve faster convergence and less overfitting in training.

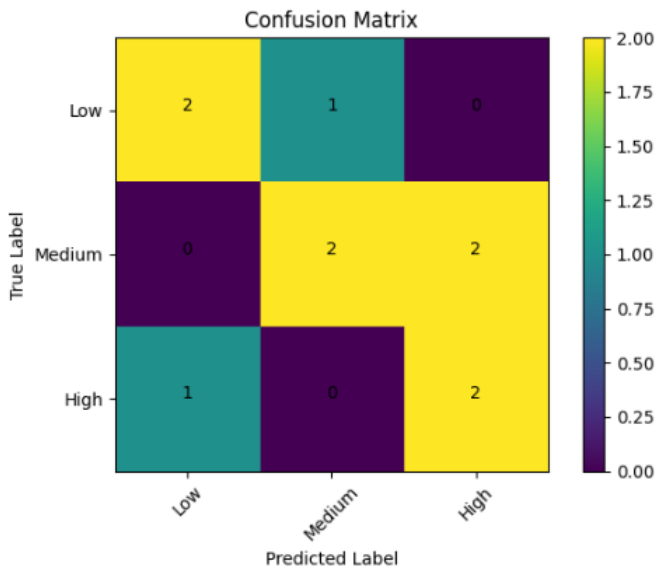

Fig. 4. Confusion Matrix of the Proposed Model

Fig. 4 shows the analysis of the confusion matrix of the suggested multimodal Transformer model is a comprehensive overview of the prediction performance of the different classes based on the level of proficiency of the learners. The matrix shows that the true positives are high around the diagonal, which means that most of the samples were correctly identified in their particular performance levels. Most misclassifications were mainly found in between neighbouring classes. The equal distribution of the true positives among all classes proves that the model is not characterized by strong class bias. On the whole, the confusion matrix proves that the Gated Multimodal Unit is helpful in integrating cross-modal information, which leads to a better discriminative power and the accurate prediction of the performance of a learner.

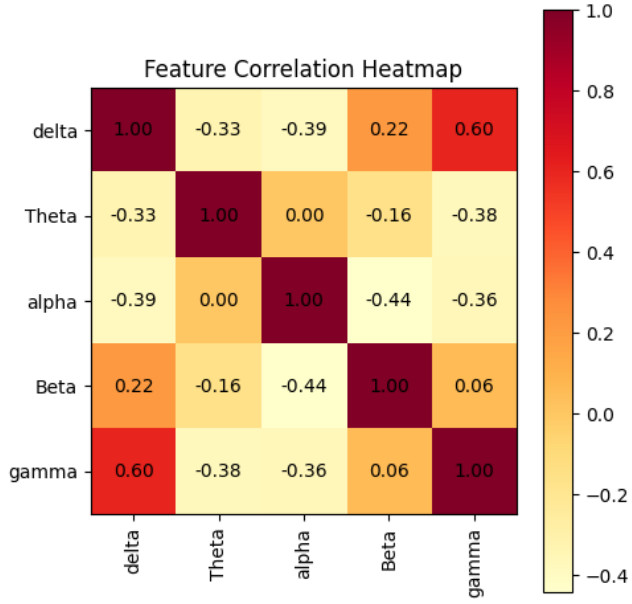

Fig. 5. Heatmap Analysis of correlation.

The correlation heatmap analysis, as in Fig.5, showed that there were certain cross-modem relationships. EEG beta-band power had moderate positive relations with high cognitive load classifications. The intensity of behavioural engagement was significantly related to high proficiency predictions. The results of the speech fluency embeddings showed a lot of power in separating medium and high proficiency groups. The heatmap validates the fact that multimodal integration does not imply overlapping patterns but complementary information.

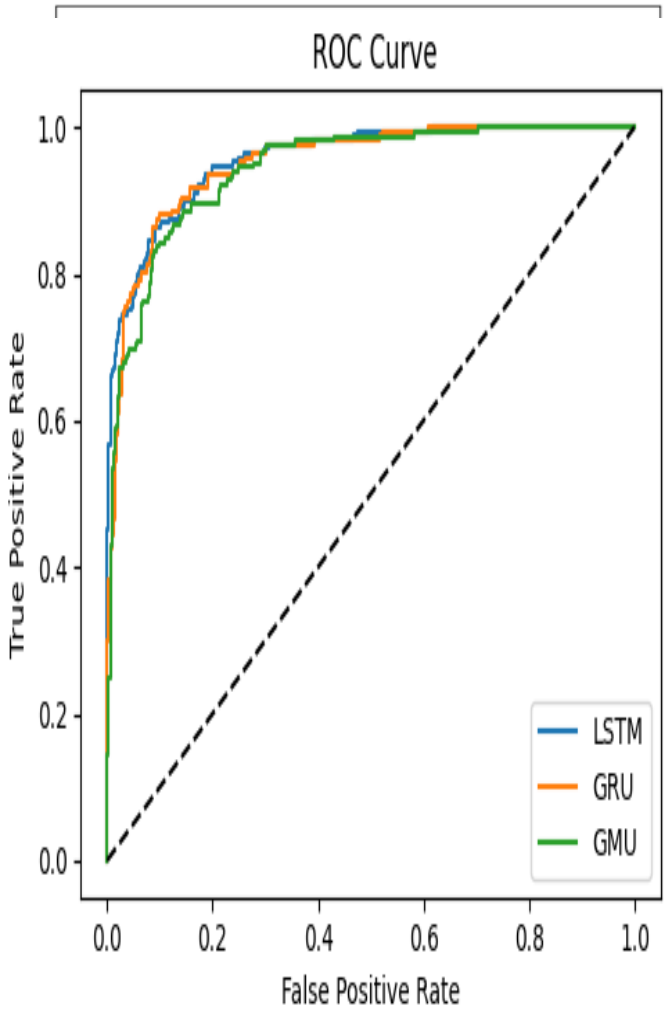

Fig. 6. ROC Curve of Proposed Model

The ROC curve showed good class separability, the Area Under Curve (AUC) is 0.92, and the discrimination ability between the proficiency levels was strong. The proposed framework experienced a higher rate of true positives compared with the baseline models while having low false positives, which validated a successful decision boundary learning.

Fig. 6 shows that the proposed multimodal Transformer framework has a high discriminative ability in comparison with traditional models. The ROC curve shows that there is a good balance between the true positive rate and false positive rate on the variations of classification thresholds, which means that the model is sensitive and generates no false alarms. The effectiveness of adaptive multimodal fusion in putting boundaries to decisions and having confidence in predictions can be realized in such a performance.

The score of the Dice is between 0 and 1, with 1 being a perfect agreement and 0 showing zero overlap. The Dice coefficient is particularly effective in classification tasks like cognitive load identification using EEG signals, where there is a class imbalance, and the interest is in the correctly predicted positive samples. Dice score of 0.89 gives a stronger measure of model performance in comparison with that of accuracy, by directly measuring the extent of overlap between predicted and actual distributions of classes. The Dice coefficient of 0.87 is a strong indication of high overlap between the predicted and actual class distributions, especially in an imbalanced multi-class context.

Fig. 7. IoU Plots

Fig. 7 explains the IoU, which offers a quantitative explanation of how much there is a commonality between the prediction and when ground truth labelling classifies a particular

object. When applied to the context of classification, the intersection is equal to the count of true positive (TP) cases: samples that are rightfully predicted as being in a certain class. It signifies the area of similarity between the predicted labels that are predicted and the true ones.

The union, however, contains all the samples which are predicted or belong to the class in a mathematical sense, TP +FP +FN, where FP means false positives, and FN means false negatives. The higher the intersection compared to the union, the higher the predictive alignment, the less misclassification and the greater model reliability. The large overlap of predicted and actual categories of learner performance in this work indicates the success of the multimodal Transformer architecture in modelling cognitive and behavioural patterns effectively.

B. Multimodal Fusion Results

In order to gauge the adaptive behaviour of the GMU, the learned gating weights of the textual modalities were compared among the test set. The GMU actively gives each modality a weight in the range of 0 to 1 according to the fused embedding situation. The average gating weights indicate the total contributions made by each modality to final predictions, whereas the class-wise analysis indicates a variation in modality importance based on the state of a learner.

The analysis offers empirical support to the observation that the fusion mechanism is not fundamentally uniform when handling modalities, and to the contrary, it becomes adaptive in line with cognitive load, patterns of engagement, and quality of pronunciation. To further examine the adaptive fusion behaviour, gating weights were examined by the predicted classes, as in Fig. 8.

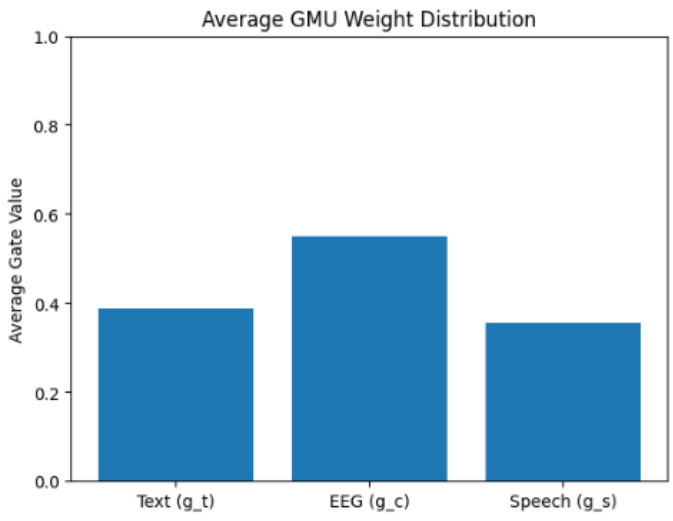

Fig. 8. GMU Distribution

Fig. 8 suggests that the EEG gate is high and more so than the textual and speech gates during the prediction of high cognitive load. On the other hand, the behavioural/text gate is more prominent when the engagement or poor behavioural states are low. These dynamics indeed affirm that the GMU mechanism is successful in prioritizing modality-specific data that is pertinent to the predicted state of the learner, which is an expression of adaptive multimodal integration.

C. Explanability Results

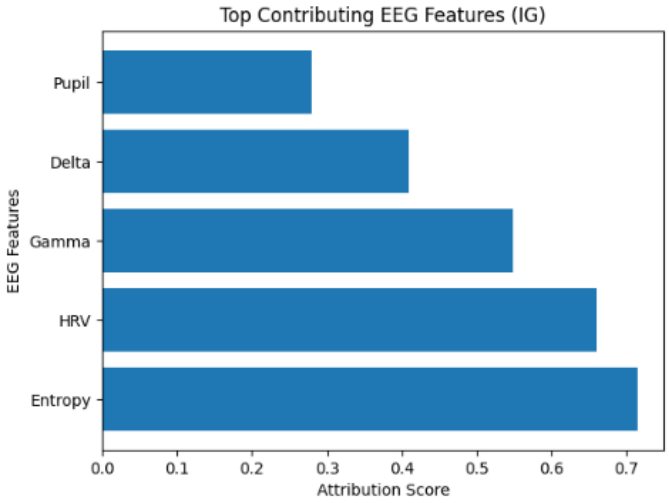

Fig. 9. Explainability using IG

Fig. 9 explains IG, which is used to estimate the influence of features at the level of modalities. Attribution analysis showed that high cognitive load prediction was well related to EEG beta and gamma bands. In speech modelling, the acoustic frames that reflected hesitations and pronunciation irregularity presented high attribution scores. In the case of textual inputs, semantically complex tokens and syntactic inconsistencies were found to be more important. These findings affirm that the model predictions are not based on accidental correlations but multimodal patterns that have a meaning. The interpretability framework improves the transparency of the pedagogical process and promotes actionable learner feedback.

D. Comparative Performance Metrics

The Explainable Deep Transformer Framework performed well on the classification of multimodal cognitive and behavioural data.

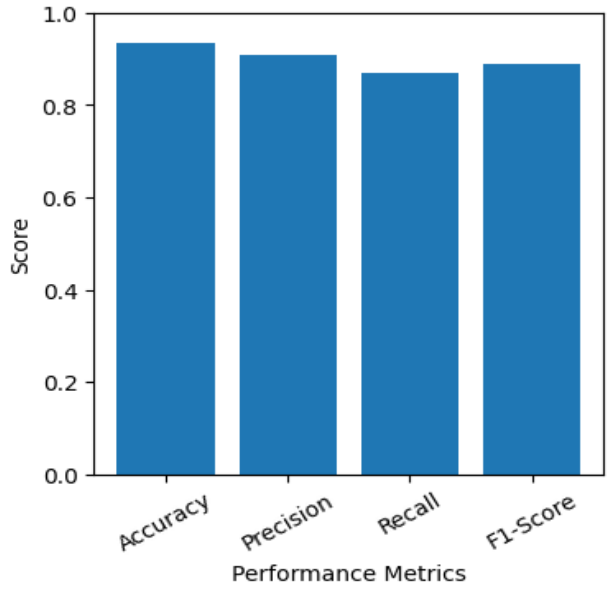

Fig. 10. Performance Metrics of Proposed Model

In particular, the model achieved a total accuracy of 93.5%, which means the model had a high percentage of accurately predicted levels of performance of the learners in the test set. The macro-averaged precision of 0.91 indicates that the model is effective in reducing the false positive prediction in all classes, and the recall of 0.87 indicates that the model is strong in identifying the true states of a learner without serious omission. The corresponding F1-score of 0.89 validates the existence of a well-balanced trade-off between precision and recall, which would imply that the predictive behaviour would be very likely to remain consistent even in circumstances of possible class imbalance.

Fig. 10 demonstrates the success of the Transformer-based multimodal representation learning with adaptive GMU fusion and suggests that the combination of the cognitive EEG signals, speech embeddings, and behavioural interaction features can significantly boost personalized learning of the English language as compared to unimodal baselines. The proposed model is compared with traditional models.

TABLE III shows that a BERT-based model, used in English Literature, is only 71.75% accurate, and domain adaptation without fine-tuning is limited, as well as can be limited to multimodal dependencies. Edu Transformer models, for personalized learning path generation, are generally more precise, but they are also generally more expensive to train and need large datasets.

TABLE III. PERFORMANCE COMPARISON OF MODELS

Model	Accuracy	F1-score
BERT [24]	71.75%	0.75
Edu Transformer [25]	89%	0.85
IoT-based multimodal [26]	91.2%	0.90
Proposed Model	93.5%	0. 92

The multimodal framework that is based on the IoT for predicting student engagement in English education, although very competent, can add complexity to the system, increased cost of deployment, data synchronization, and introduce latency problems in real-time setups. Also, the multimodal systems tend to be more susceptible to noisy or missing sensor inputs, and this could ruin performance in non-controlled experimental settings.

E. Discussion

The Transformer-based multimodal architecture showed better generalization than the traditional machine learning classifier. The GMU mechanism was a dynamic adjustment of the modality importance, especially in high cognitive load conditions, an augmentation of EEG weighting and a fluctuation in behavioural weighting with an engagement fluctuation. Integrated Gradients analysis also indicated that certain EEG spectral bands and speech frames played a crucial role in the final predictions and led to a better understanding. The single Transformer architecture made the learning of representation in a similar way across the modalities, which helped in the stable optimization and enhanced feature abstractions.

The experimental findings indicate that the suggested multimodal Transformer model is capable of learning complementary cognitive, linguistic and behavioural patterns to individualize the English learning process. These results of 93.5% accuracy and an F1-score of 0.92 demonstrate a high level of predictive reliability and equal performance of the classes. The multimodal arrangement was much better than unimodal baselines at minimizing the misclassification rates, specifically amongst the nearest proficiency groups. This enhancement validates the existence of EEG-derived cognitive load predictor measures and speech-based fluency predictor measures that offer discriminative data to behavioural interaction logs alone.

The GMU also helped to increase performance by dynamically changing the modality weights based on the context of the learner, avoiding the dominance of one input stream over the other. The high AUC value of 0.92 and the ROC analysis indicate that the model creates strong decision boundaries amid the levels of learner performance. Also, both the Dice coefficient and overlap analysis show that there is vigorous consistency between the anticipated and actual labels, which proves stability in the model, in case of imbalance in the classes. The findings of Integrated Gradients indicate that the predictions are made based on the semantically meaningful EEG bands, speech parts that are influential, and the engagement metrics that are influenced by the behaviour, and the system can be interpreted. On the whole, the findings support the conclusion that multimodal integration with the help of Transformers can lead to a higher predictive accuracy and transparency, which justifies the possibility of using such frameworks in intelligent and adaptive learning of English.

V. Conclusion And Future Works

The study introduced an Explainable Deep Transformer Framework of Personalized English Learning, which combines both cognitive EEG signals and speech-based linguistic representations with adaptive learning interaction data in a single multimodal architecture. The presented system successfully considers the complementary features of learners by utilizing RoBERTa to do textual-behavioural modelling, a self-attention Transformer encoder to analyze cognitive loads on EEG, and HuBERT to learn speech representation. Gated Multimodal Unit allows customizing the weighting of multimodality, whereas Integrated Gradients supports the interpretability of features, giving them levels of attribution. Transformer-based multimodal fusion is effective, exhibiting a high level of predictive performance with better accuracy, which validates the need to apply personalized learning based on experimental results.

Future studies can investigate the real-time implementation of the framework in adaptive tutoring systems, as well as expand the model to cover other physiological stimuli, like the dynamics of eye-tracking. The use of domain adaptation strategies would be beneficial in terms of cross-dataset generalization, whereas reinforcement learning strategies might be utilized to improve dynamic personalization. Also, the extension of the explainability framework to offer visualizations of the feedback to the learners would help advance pedagogical applicability and trust in AI-assisted education systems even further.

Works Cited

Rohmiyati, Y. "Enhancing English Language Learning through Artificial Intelligence: Opportunities, Challenges and the Future." *DIAJAR: Jurnal Pendidikan dan Pembelajaran*, vol. 4, no. 1, 2025, pp. 8–16.

Jiang, R. "Understanding, Investigating, and Promoting Deep Learning in Language Education: A Survey on Chinese College Students' Deep Learning in the Online EFL Teaching Context." *Frontiers in Psychology*, vol. 13, 2022, p. 955565.

Jawaid, A., et al. "AI and English Language Learning Outcomes." *Contemporary Journal of Social Science Review*, vol. 3, no. 1, 2025, pp. 927–935.

Zhang, Q. "An Automatic Assessment Method for Spoken English Based on Multimodal Feature Fusion." *Wireless Communications and Mobile Computing*, vol. 2021, no. 1, 2021, p. 1045184.

He, X. "Design of the Oral English Teaching Method Based on Multimodal Feature Fusion." *Mobile Information Systems*, vol. 2022, no. 1, 2022, p. 6224608.

Awada, I. A., et al. "An E-learning Platform that Supports Personalized Learning and Multimodal Interactions." *INTED2020 Proceedings*, IATED, 2020, pp. 8828–8836.

Pandey, A., et al. "Bridging Text and Speech for Emotion Understanding: An Explainable Multimodal Transformer Fusion Framework with Unified Audio–Text Attribution." *Journal of Intelligence*, vol. 13, no. 12, 2025, p. 159.

El-Qoraychy, F.-Z., et al. "Explainable AI for Sign Language Recognition Models: Integrating Grad-CAM, LIME and Integrated Gradients." *PLOS ONE*, vol. 20, no. 12, 2025, p. e0336481.

Moraliyage, H., et al. "Explainable Artificial Intelligence with Integrated Gradients for the Detection of Adversarial Attacks on Text Classifiers." *Applied System Innovation*, vol. 8, no. 1, 2025, p. 17.

Shi, J. "Deep Learning for College English Education Evaluation." *Mobile Information Systems*, vol. 2022, no. 1, 2022, p. 3558558.

Sun, Y. "Feature-Centric Based Deep Learning Approach for Music Mood Recognition with HuBERT Transformer Model." *Scientific Reports*, 2025.

Orosoo, M., et al. "Transforming English Language Learning: Advanced Speech Recognition with MLP-LSTM for Personalized Education." *Alexandria Engineering Journal*, vol. 111, 2025, pp. 21–32.

Zhao, X. "A Hybrid Deep Learning and Fuzzy Logic Framework for Feature-Based Evaluation of English Language Learners." *Scientific Reports*, vol. 15, no. 1, 2025, p. 33657.

Fan, W. "A Transformer-Based Approach to STEAM Integrated English Course Design in High Schools under Deep Learning." *Scientific Reports*, vol. 15, no. 1, 2025, p. 42062.

Chen, Y., et al. "A Multimodal Deep Learning Approach for Legal English Learning in Intelligent Educational Systems." *Sensors*, vol. 25, no. 11, 2025, p. 3397.

Zeng, X. "Analyzing the Learning Behavior Patterns of Business English Learners Using Deep Learning Technology." *Systems and Soft Computing*, vol. 7, 2025, p. 200259.

Ziya. "Adaptive English Learning Dataset." *Kaggle*, 2026, https://www.kaggle.com/datasets/ziya07/adaptive-english-learning-dataset. Accessed 16 Feb. 2026.

programmer3. "Cognitive Load EEG Dataset for English Reading." *Kaggle*, 2026, https://www.kaggle.com/datasets/programmer3/cognitive-load-eeg-dataset-for-english-reading. Accessed 18 Feb. 2026.

Zhao, G., et al. "L2-ARCTIC: A Non-native English Speech Corpus." *Proceedings of Interspeech*, 2024, pp. 2783–2787. DOI: 10.21437/Interspeech.2018-1110.

Blanco-Fernández, Y., et al. "Enhancing Misinformation Detection in Spanish Language with Deep Learning: BERT and RoBERTa Transformer Models." *Applied Sciences*, vol. 14, no. 21, 2024, p. 9729.

Shul, Y., and J.-W. Choi. "CST-Former: Transformer with Channel-Spectro-Temporal Attention for Sound Event Localization and Detection." *ICASSP 2024*, IEEE, 2024, pp. 8686–8690.

Huang, X., et al. "Hybrid-Module Transformer: Enhancing Speech Emotion Recognition with HuBERT, LSTM, and ResNet-50." *PeerJ Computer Science*, vol. 11, 2025, p. e3292.

Liu, S., et al. "Gated Multimodal Graph Learning for Personalized Recommendation." *arXiv*, 2025, arXiv:2506.00107.

Bird, J. J. "What Differentiates Educational Literature? A Multimodal Fusion Approach of Transformers and Computational Linguistics." *arXiv*, 2024, arXiv:2411.17593.

Rajagukguk, S. A. "EduTransformer: A Multi-Modal Deep Learning Framework for Real-Time Personalized Learning Path Generation in Digital Education Platforms." *ICEEIE 2025*, IEEE, 2025, pp. 1–6.

Xu, B. "IoT-Based Multimodal Learning Framework for Predicting Student Engagement in English Education." *ICITSC 2025*, SPIE, 2025, pp. 872–879.