Multimodal Technologies and Interaction

MDPI

Article

Parallel Bilingual Datasets: A Multimodal Deep Learning Framework for Proficiency and Style Classification

Padmavathi Kesavan [1], Miranda Lakshmi Travis [1], Martin Aruldoss [2] and Martin Wynn [3,*]

[1] PG and Research Department of Computer Science and AI, St. Joseph's College of Arts & Science (Autonomous), Annamalai University, Cuddalore 607001, India; padmavathi@sjctnc.edu.in (P.K.); miranda@sjctnc.edu.in (M.L.T.)
[2] Department of Computer Science, Central University of Tamil Nadu, Thiruvarur 610005, India; martin@cutn.ac.in
[3] School of Business, Computing and Social Sciences, University of Gloucestershire, Cheltenham GL50 2RH, UK
* Correspondence: mwynn@glos.ac.uk

Abstract

This study presents a multimodal deep learning framework for automatic proficiency and style classification of parallel Bilingual Tamil–Hindi learner data. The proposed system employs a dual-headed neural architecture to simultaneously predict proficiency levels (Basic, Advanced) and stylistic categories (Formal, Literary) using shared feature representations. A curated dataset of bilingual text samples is utilized, along with synthetic speech generated through text-to-speech (TTS) to enable controlled multimodal experimentation. Five deep learning architectures are evaluated under text-only, audio-only, and learnable fusion settings. Experimental findings indicate that text-based models consistently achieve strong performance in both proficiency and style classification tasks. In contrast, the audio-only model demonstrates limited effectiveness, highlighting the constraints of synthetic acoustic features in capturing meaningful linguistic information. The fusion models provide only marginal improvements over text-based approaches, suggesting that textual representations play a dominant role in proficiency and stylistic classification within controlled datasets. These results emphasize the importance of linguistic features over acoustic signals for automated language assessment in low-resource settings. The proposed framework provides a scalable and reproducible approach and offers a foundation for future work incorporating real speech data and more diverse linguistic inputs.

Keywords: multimodal learning; language proficiency classification; style classification; deep learning; Tamil–Hindi dataset

Check for updates

Academic Editor: Stephan Schlögl

Received: 30 March 2026
Revised: 25 April 2026
Accepted: 27 April 2026
Published: 30 April 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

In recent years, the rapid advancement of artificial intelligence and deep learning has significantly transformed automated language assessment systems, enabling scalable, objective, and data-driven evaluation of linguistic proficiency. This transformation is particularly important in multilingual environments such as India, where languages like Tamil exhibit rich morphological structures, agglutinative properties, and diverse stylistic variations. These characteristics make traditional rule-based and statistical approaches inadequate, thereby necessitating robust deep learning-based frameworks for accurate classification and assessment [1–3].

Deep learning models have demonstrated remarkable success in text classification tasks. Convolutional Neural Networks (CNNs) effectively capture local semantic patterns and n-gram features [3], while Long Short-Term Memory (LSTM) networks model long-range dependencies in sequential data. Furthermore, Bidirectional LSTM (BiLSTM) architectures enhance contextual understanding by processing sequences in both forward and backward directions [4–8]. In addition, hybrid architectures such as CNN + LSTM and CNN + BiLSTM [9–12] have been shown to outperform standalone models by combining local feature extraction with sequential learning capabilities. Recent studies [13,14] demonstrate that hybrid architectures improve classification accuracy by combining local and sequential features [13–20]. The use of subword-level embeddings such as FastText further improves performance for morphologically rich and low-resource languages by capturing character-level and contextual information [5].

Despite these advancements, text-only models fail to capture paralinguistic features such as pronunciation, intonation, and speech fluency, which are critical for comprehensive language proficiency and stylistic evaluation. Speech-based models, particularly those leveraging deep convolutional architectures and self-supervised learning techniques, have demonstrated strong capability in extracting acoustic features for language-related tasks [21–24]. However, standalone audio models often exhibit lower performance due to noise sensitivity and limited contextual representation.

To overcome these limitations, multimodal learning has emerged as a powerful paradigm that integrates complementary information from multiple modalities such as text and audio. Fusion techniques, including early fusion, late fusion, and learnable fusion, enable effective combination of heterogeneous features. Among these, learnable fusion approaches allow adaptive weighting of modalities, thereby improving robustness and classification accuracy [25–28]. These findings support the use of CNN–BiLSTM in this study for capturing both local and contextual linguistic features.

The motivation for this research stems from the limitations of existing systems, which predominantly rely on unimodal approaches and fail to jointly capture linguistic and acoustic characteristics. In morphologically rich languages like Tamil, both textual semantics and speech patterns play a crucial role in determining proficiency and stylistic variation. Therefore, this study aims to develop a unified multimodal deep learning framework that integrates text and audio features within a dual-headed multi-task learning architecture to simultaneously classify proficiency and style.

Following this brief introduction, Section 2 sets out the research method, comprising a review of relevant literature and a controlled experiment. Section 3 then presents an overview of relevant literature and sets out three research questions that the article addresses. Section 4 sets out the main elements of the experiment design, and Section 5 then directly addresses the research questions. A multimodal framework is proposed for joint classification of language proficiency and writing style using Tamil text and corresponding audio features. Then, multiple deep learning architectures, including CNN, LSTM, BiLSTM, CNN + LSTM, and CNN + BiLSTM, are systematically evaluated under a unified setup, and an audio-based convolutional model is incorporated to extract acoustic representations. A learnable fusion mechanism is introduced to effectively combine predictions from text and audio modalities, and an extensive experimental analysis is conducted to compare unimodal and multimodal approaches, demonstrating the effectiveness of the proposed framework. Section 6 then discusses some emergent issues and Section 7 concludes the study.

2. Materials and Methods

This study consisted of two main research phases. Phase 1 comprised an integrative literature review, underpinned by an interpretivist philosophy and a qualitative approach. This combination is particularly appropriate for exploratory research when the researchers are looking to find an explanation of the phenomenon under study [29]. Snyder [30] (p. 335) notes that "for newly emerging topics, the purpose [of an integrative review] is rather to create initial or preliminary conceptualizations and theoretical models, rather than review old models. This type of review often requires a more creative collection of data, as the purpose is usually not to cover all articles ever published on the topic but rather to combine perspectives and insights from different fields or research traditions".

The review systematically analyzes existing methods across three dimensions: (i) unimodal text classification, (ii) speech-based modeling, and (iii) multimodal learning frameworks. By synthesizing findings from these domains, the study identifies key limitations in current approaches, particularly the lack of multi-task multimodal models for low-resource languages. This integrative perspective provides the foundation for the proposed framework, which aims to bridge the gap between textual and acoustic analysis for comprehensive language assessment.

Phase 2 involved a controlled experiment to develop and evaluate a deep learning framework for proficiency and style classification using parallel Bilingual Tamil–Hindi data. The dataset consists of 2229 manually constructed sentence pairs, where each Tamil sentence is aligned with its corresponding Hindi translation representing the same semantic content. Unlike datasets collected from uncontrolled sources such as social media or open corpora, the sentences were manually authored to maintain linguistic clarity, consistent annotation, and controlled variation in proficiency and style. This design enables focused evaluation of linguistic features while minimizing noise and inconsistencies. Each record contains a Tamil sentence along with its corresponding Hindi translation, and is annotated with proficiency (Basic/Advanced) and style (Formal/Literary) labels. The dataset underwent systematic preprocessing, including duplicate removal, label consistency verification, and normalization of textual content to ensure high annotation quality.

The class distribution was maintained to ensure adequate representation across proficiency levels, while stylistic categories reflect realistic but controlled variation shown in Table 1. Although the dataset is synthetically constructed, it is designed to capture structured linguistic patterns suitable for evaluating classification models under controlled conditions. This approach enables the analysis of model performance based on linguistic characteristics, without confounding factors such as topic variability or uncontrolled stylistic noise.

Table 1. Distribution of Proficiency and Style.

Proficiency	Style	Count
Basic	Formal	558
Basic	Literary	564
Advanced	Formal	540
Advanced	Literary	567
Total		2229

Text Representation: Each sentence in the dataset is annotated along two linguistic dimensions: proficiency level and stylistic form. The proficiency level captures the linguistic complexity of the sentence. Sentences labelled as Basic typically contain simple grammatical constructions, limited vocabulary, and short sentence structures commonly produced by beginner learners. In contrast, Advanced sentences exhibit richer vocabulary,

more complex syntactic structures, and greater semantic depth, reflecting higher levels of language proficiency. The stylistic form captures the communicative style of the sentence. Formal sentences represent standard language used in educational, professional, or instructional contexts, whereas Literary sentences contain expressive or descriptive elements often associated with narrative or creative writing styles. To illustrate these annotations, representative examples from the dataset are provided in both Tamil and Hindi in Appendix A. The dataset is constructed as a parallel bilingual corpus, where each sample consists of a Tamil sentence and its corresponding Hindi translation conveying the same semantic content. The proficiency and style labels are assigned at the sentence level and are shared across both language representations. This dataset does not include code-switched or mixed-language sentences.

The comparison between traditional language evaluation and the proposed AI-based framework is illustrated in Figure 1. In the traditional approach, a learner's sentence is manually evaluated by human experts based on grammatical correctness, vocabulary usage, and stylistic appropriateness. This process is inherently subjective, time-consuming, and difficult to scale for large datasets.

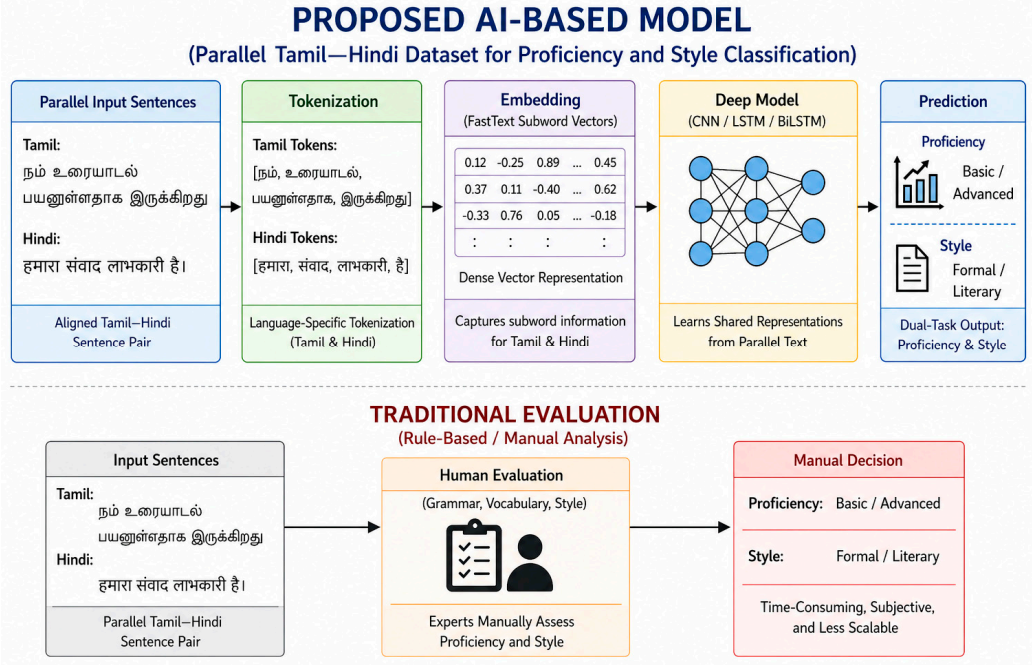


Figure 1. Comparison between traditional evaluation and the proposed AI-based Model.

In contrast, the proposed AI-based system follows a fully automated pipeline. Each input consists of a parallel Bilingual Tamil–Hindi sentence pair, such as the Tamil sentence "நம் உரையாடல் பயனுள்ளதாக இருக்கிறது". and its corresponding Hindi translation "हमारा संवाद लाभकारी है।" (as shown in Figure 1: English translation: "Our conversation is useful"). The sentences are first tokenized into individual linguistic units for both languages. These tokens are then transformed into dense vector representations using FastText embeddings, which capture subword-level and semantic information and effectively handle morphologically rich languages. The embedded representations are subsequently processed by deep learning models such as CNN, LSTM, or BiLSTM to learn structural and linguistic patterns from the combined text representation. Finally, the model simultaneously predicts the proficiency level (Basic/Advanced) and stylistic category (Formal/Literary) of the input.

This comparison highlights the key advantages of the proposed system, including scalability, consistency, and the ability to perform real-time classification without human intervention.

Synthetic Audio Generation for Multimodal Learning: To facilitate multimodal experimentation, synthetic speech signals were generated for all 2229 text samples using a text-to-speech (TTS) pipeline. This approach enables the creation of a parallel audio modality without relying on human speakers or real-world recordings. Synthetic speech offers several advantages, including consistent pronunciation, controlled speaking style, elimination of speaker bias, and scalability for low-resource language research.

From the generated audio signals, Mel-Frequency Cepstral Coefficients (MFCCs) were extracted to represent phonetic and spectral characteristics of speech. Each audio sample was transformed into a fixed-dimensional feature representation, resulting in a corresponding audio dataset aligned one-to-one with the textual samples. These audio features were used exclusively for modelling and evaluation purposes and were not transcribed back into text.

Multimodal Alignment: The final dataset therefore consists of:

- 2229 text samples represented as tokenized sequences for textual modeling;
- 2229 MFCC-based acoustic feature representations extracted from synthetically generated speech for audio modeling.

Both modalities share identical labels, enabling controlled evaluation of text-only, audio-only, and multimodal fusion architectures shown in Figure 2. This design allows for a systematic investigation of the contribution of synthetic audio features to proficiency classification while maintaining semantic consistency across modalities.

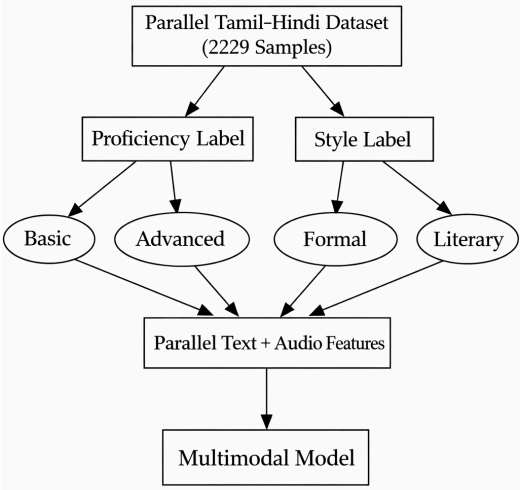


Figure 2. Overview of the Tamil–Hindi dataset and annotation pipeline.

By explicitly constructing both text and audio data, the dataset supports reproducible experimentation and serves as a reliable benchmark for evaluating multimodal deep learning models in educational language assessment for morphologically rich languages.

All experiments were conducted using Python in the Google Colab environment. Deep learning models were implemented using TensorFlow (version 2.19.0) and Keras (version 3.13.2). Text preprocessing, including tokenization and padding, was performed using the Keras Tokenizer. FastText embeddings were utilized for text representation. Audio features were extracted using Librosa (version 0.11.0), and Mel-Frequency Cepstral Coefficients (MFCCs) were computed for acoustic representation. Synthetic speech signals were generated using a text-to-speech (TTS) system available in the Google Colab environment.

3. Relevant Literature

Deep learning has become the dominant paradigm for language classification and assessment tasks, particularly in low-resource and morphologically rich languages. Early approaches relied on traditional machine learning techniques such as Support Vector Machines (SVMs), which were limited in capturing complex linguistic patterns [12]. With the emergence of deep neural networks, significant improvements have been achieved in text classification tasks.

CNN-based architectures have been widely used for text classification due to their ability to capture local features and semantic patterns. Kim [3] demonstrated the effectiveness of CNNs for sentence classification tasks. Subsequent studies extended CNN architectures for multilingual and domain-specific applications, showing improved performance in capturing contextual information [1,2,6]. A range of models are discussed in the literature.

LSTM-based Models are designed to capture long-term dependencies in sequential data and have been successfully applied to language modeling and classification tasks. The foundational work by Hochreiter and Schmidhuber [4] introduced LSTM as a solution to the vanishing gradient problem. Later studies demonstrated the effectiveness of LSTM models in handling sequential text data for sentiment and language classification tasks [7,8,15]. Bidirectional LSTM models extend LSTM by processing sequences in both forward and backward directions, enabling better contextual representation. BiLSTM has been particularly effective for languages with complex syntax and word order variations. Studies have shown that BiLSTM significantly improves classification performance in Tamil and other Indian languages [9–11].

Hybrid architectures combining CNN and LSTM have gained attention due to their ability to capture both local and sequential features. CNN layers extract local patterns, while LSTM layers model temporal dependencies. The effectiveness of CNN-LSTM models in language classification tasks demonstrated in [6,7]. Recent studies further confirmed the superiority of hybrid architectures over standalone models [13,14]. CNN–BiLSTM models represent an advanced hybrid approach that combines convolutional feature extraction with bidirectional sequence modeling. These models have shown superior performance in complex classification tasks due to their ability to capture both contextual and sequential information effectively. Recent works have highlighted their effectiveness in multilingual and low-resource scenarios [17–20].

Speech and Audio-based Models have gained significant attention with the advancement of deep learning [21] introduced self-supervised learning techniques for speech representation, enabling improved performance in speech recognition tasks. Other studies have demonstrated the effectiveness of deep neural networks in extracting acoustic features for classification tasks [22,24]. However, audio-only models often face challenges related to noise and variability in speech signals.

CNN-Based Audio Models (2D Spectrogram Learning): Recent advancements in deep learning have enabled the effective application of two-dimensional Convolutional Neural Networks (2D CNNs) for audio classification tasks [31]. Unlike traditional approaches that rely on handcrafted features, modern audio models transform raw audio signals into time–frequency representations such as spectrograms or Mel-spectrograms, which can be treated as images and processed using CNN architectures [32].

Spectrogram-based representations preserve both temporal and frequency-domain information, making them highly suitable for capturing acoustic patterns such as pitch, tone, and energy variations. Studies have shown that converting audio signals into spectrogram images significantly improves classification performance compared to raw waveform in-

puts [33]. In this approach, CNNs learn hierarchical feature representations directly from these 2D inputs, similar to image recognition tasks.

Several works have demonstrated the effectiveness of 2D CNN architectures for audio classification [31] proposed a frequency-based CNN model that extracts discriminative features from spectrogram representations, improving classification accuracy in speech-related tasks [34]. Similarly, Ashurov et al. [32] employed multiple pre-trained CNN architectures such as ResNet and DenseNet on spectrogram images, achieving high accuracy in environmental sound classification tasks.

Furthermore, research by Cheng et al. [33] demonstrated the use of CNN-based models with spectrogram inputs for vehicle sound classification, achieving high accuracy in real-world noisy environments. Also, Log-Mel spectrogram representations improve CNN-based audio classification by capturing perceptually relevant frequency features [35]. Other studies have also explored CNN-based architectures for environmental sound classification and acoustic scene analysis, confirming that CNNs outperform traditional machine learning approaches when applied to spectrogram features [36,37].

In addition, modern architectures such as YAMNet and transfer learning-based CNN models have further improved audio classification by leveraging large-scale pre-trained networks trained on datasets like AudioSet [38]. These models utilize deep convolutional layers to automatically extract robust acoustic features, reducing the need for manual feature engineering.

Overall, 2D CNN-based audio models have become a standard approach for speech and sound classification tasks due to their ability to effectively model complex acoustic patterns. In the context of this study, the use of a 2D CNN architecture on audio features (MFCC/spectrogram-like inputs) aligns with established methodologies and enables the extraction of meaningful acoustic representations for proficiency and style classification.

Recent advancements in artificial intelligence have demonstrated the effectiveness of attention-based deep learning models across diverse domains, including real-time object detection and tracking. For instance, attention-enhanced YOLO-based frameworks have been successfully applied in complex visual environments, highlighting the importance of feature representation and model adaptability. Similarly, deep learning approaches for text style transfer have gained significant attention, focusing on modifying stylistic attributes while preserving semantic content. These developments underline the growing importance of representation learning and style modeling, which are also central to the proposed framework [39,40].

Multimodal learning integrates information from multiple modalities to improve model performance. Baltrušaitis et al. [25] provided a comprehensive survey on multimodal machine learning, highlighting the importance of fusion strategies. Ramachandram and Taylor [26] discussed deep multimodal learning techniques, while Poria et al. [27] and Kumar et al. [28] demonstrated the effectiveness of multimodal fusion in improving classification accuracy. These studies emphasize that combining text and audio modalities leads to more robust and reliable language assessment systems.

In summary, existing research demonstrates that different deep learning architectures contribute uniquely to language classification tasks. Convolutional Neural Networks (CNNs) are particularly effective in capturing local lexical and stylistic patterns, making them suitable for writing style classification [3,16]. In contrast, sequential models such as Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM) are better at modeling contextual dependencies and long-range relationships within text, which are essential for proficiency assessment [4,9,15].

Hybrid architectures, including CNN–LSTM and CNN–BiLSTM, combine the strengths of both convolutional and recurrent networks, enabling the extraction of both local and

global linguistic features. Prior studies have shown that such hybrid models consistently outperform standalone architectures in complex language processing tasks [6,7,17]. This makes them particularly suitable for multilingual and morphologically rich language settings.

On the other hand, audio-based models, especially those using 2D CNNs on spectrogram or MFCC representations, have demonstrated the ability to capture acoustic features such as intonation, rhythm, and pronunciation [21,22]. However, standalone audio models often exhibit limited discriminative power due to noise sensitivity and lack of contextual linguistic information, resulting in comparatively lower classification performance.

To address these limitations, multimodal learning approaches have been widely explored. Fusion strategies, particularly late fusion, have been shown to provide more stable and interpretable results by combining independent predictions from text and audio modalities [25,26]. In such frameworks, textual features typically contribute the dominant signal, while audio features act as complementary cues that enhance robustness and generalization. This aligns with recent findings that multimodal systems outperform unimodal approaches in educational and language assessment applications [27,28].

Based on the identified research gaps, the following research questions (RQs) guide this study:

RQ1: How effectively can deep learning models (both standalone and hybrid) classify parallel Bilingual Tamil–Hindi learner sentences into proficiency levels and stylistic categories using textual features?

RQ2: To what extent do acoustic features derived from synthetic speech contribute to proficiency and style classification when compared with text-based representations?

RQ3: Does multimodal fusion of textual and audio predictions improve classification performance compared to unimodal models?

Addressing these research questions enables a systematic investigation of multimodal deep learning approaches for parallel Tamil–Hindi language assessment and contributes toward the development of scalable educational language technologies for low-resource linguistic settings. These studies collectively demonstrate the effectiveness of hybrid and multimodal approaches, which motivates the use of a CNN–BiLSTM-based multimodal framework in this study.

4. Experimental Design

4.1. Proposed Multimodal Classification Model Subsection

The proposed system investigates text-based, audio-based, and multimodal language proficiency classification for parallel Bilingual Tamil–Hindi sentence pairs. Unlike traditional multimodal systems that rely on real speech and Automatic Speech Recognition (ASR), this study adopts a controlled multimodal framework in which synthetic speech signals are generated directly from text to enable reproducible experimentation. The framework is designed to classify sentence level inputs into proficiency levels and stylistic categories by integrating complementary linguistic cues derived from text and audio modalities.

For the text modality, input Tamil–Hindi sentences are first normalized, tokenized, and padded to a fixed length. Each token is mapped to a dense vector using FastText embeddings trained on the dataset, which are particularly effective for morphologically rich languages due to their subword modelling capability. The embedded sequences are then processed using multiple deep learning architectures, including CNN, LSTM, BiLSTM, CNN + LSTM, and CNN + BiLSTM. The convolutional layers extract local n-gram and stylistic features such as formality markers, while the subsequent bidirectional LSTM captures long-range contextual dependencies by modelling both past and future word se-

quences. A dropout layer is applied to reduce overfitting, followed by a fully connected layer that learns a compact shared representation for classification.

For the audio modality, speech inputs corresponding to the same sentences are transformed into two-dimensional acoustic feature representations. These features are passed through a convolutional neural network consisting of stacked Conv2D and MaxPooling layers, which effectively learn time–frequency patterns relevant to pronunciation and prosodic cues. The extracted audio features are flattened and regularized using dropout before being passed through a fully connected layer with a softmax classifier to produce class probabilities.

To combine both modalities, a fusion strategy is employed. Instead of merging raw features, the late fusion approach combines posterior probability scores from text and audio models using weighted averaging. In contrast, the learnable fusion method integrates modality-specific representations through a trainable layer, enabling the model to automatically learn optimal fusion weights during training. This fusion approach allows the text modality, shown to be more reliable for proficiency prediction, to contribute more strongly, while still incorporating complementary information from speech signals. The final prediction is obtained by selecting the class with the highest fused probability as illustrated in Figure 3.

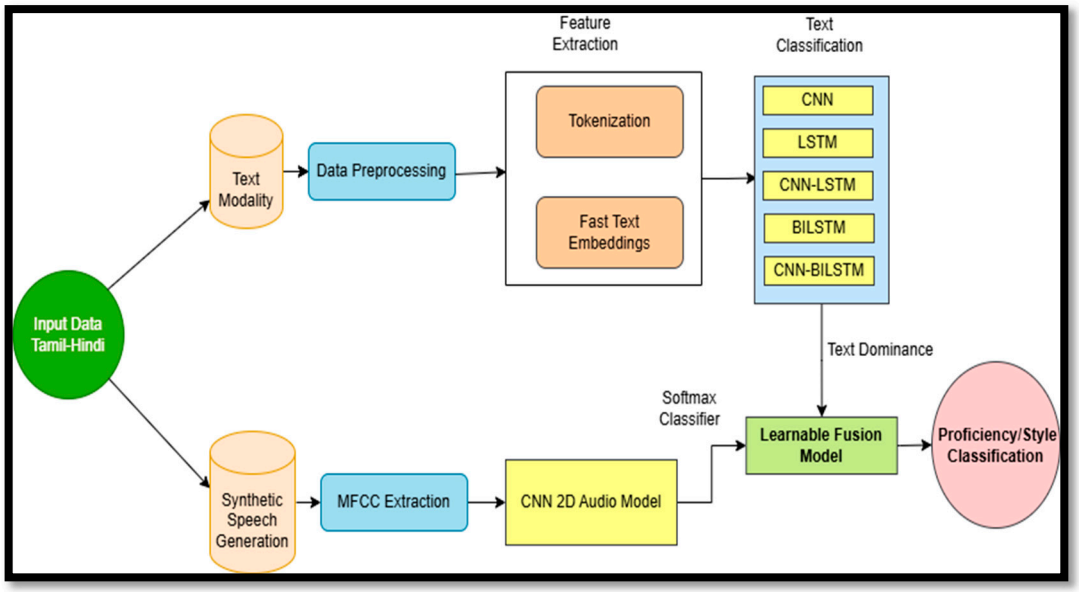


Figure 3. Dual-headed multimodal model architecture using text and audio.

The proposed modular architecture allows the text and audio models to be trained and optimized independently, while still enabling their integration through a flexible fusion mechanism. This design simplifies model development and makes the framework easily extendable to future enhancements, such as incorporating real learner speech data or supporting additional languages. By combining textual representations that capture linguistic structure with acoustic features that reflect pronunciation and speech patterns, the model provides a more comprehensive understanding of sentence-level language patterns. This integrated approach improves the robustness and reliability of classification, particularly in low-resource and parallel Bilingual Tamil–Hindi educational settings where textual cues play a dominant role, while spoken cues provide complementary information.

4.2. Data Representation

The dataset used in this study was explicitly created by the authors to support automated language proficiency assessment and tutoring applications. As described in

Section 2, the dataset consists of 2229 parallel Bilingual Tamil–Hindi sentence pairs representing commonly used day-to-day communication patterns. Each sentence is manually annotated across two linguistic dimensions: (i) proficiency level (Basic, Advanced) and (ii) stylistic form (Formal, Literary). These annotations enable the proposed framework to perform dual classification tasks, where the model simultaneously predicts the sentence-level proficiency and stylistic form.

To ensure the reliability and consistency of the annotation process, a structured labeling protocol was followed based on predefined linguistic criteria, including sentence complexity, vocabulary richness, and stylistic expression. In addition, a consistency evaluation was conducted on a subset of the dataset using Cohen's Kappa coefficient. The results indicate substantial agreement, with a score of 0.90 for both proficiency and style classification tasks. This demonstrates a high level of consistency in the annotation process.

Before model training, the dataset underwent several preprocessing steps to ensure data quality and consistency. These steps included text normalization, removal of duplicate entries, tokenization, and label verification. In particular, duplicate sentence detection was performed using string matching techniques to ensure that no identical sentences were repeated within the dataset shown in Figure 4. This step was necessary to prevent data leakage between training and validation sets and to ensure reliable evaluation of model performance.

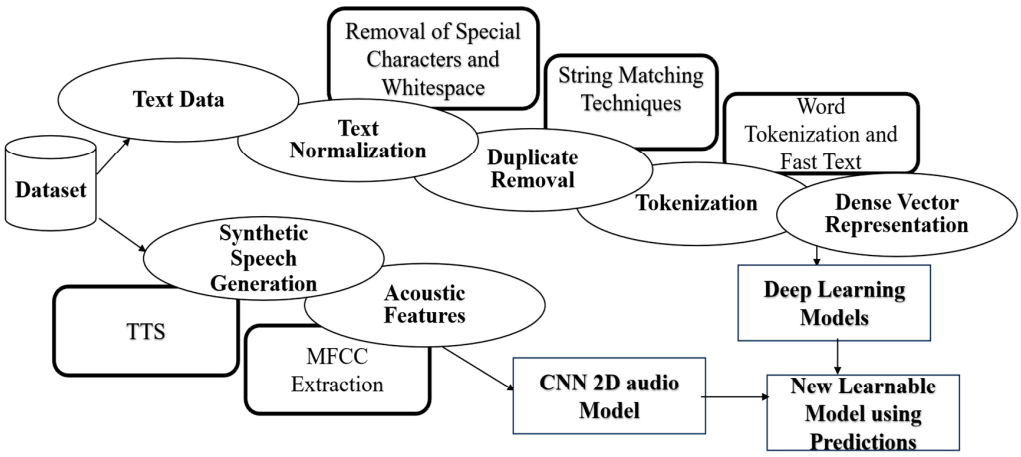


Figure 4. Dataset Representation and Preprocessing.

The dataset does not contain text or speech collected from human subjects. Instead, the sentences were synthetically authored by the researchers to represent structured sentence patterns commonly encountered in classroom exercises and language training materials. This controlled dataset design ensures balanced linguistic complexity and avoids noise introduced by uncontrolled user-generated data sources.

For computational processing, the text data were converted into tokenized sequences and padded to a fixed length to maintain consistent input dimensions across deep learning models. Each token was then mapped to a dense vector representation using FastText embeddings trained on the dataset, which capture subword-level information and are particularly suitable for morphologically rich languages such as Tamil and Hindi. This representation allows the neural network models to learn both lexical patterns and contextual relationships within parallel Bilingual Tamil–Hindi sentences, supporting accurate classification of proficiency and stylistic features. To enable multimodal experimentation, synthetic speech signals were generated from the text samples using text-to-speech techniques, and corresponding acoustic features were extracted. This design allows controlled evalu-

ation of text-only, audio-only, and multimodal learning architectures without the ethical and privacy concerns associated with real learner speech data.

A comprehensive search was conducted across major open repositories, including Kaggle, Hugging Face Datasets, and GitHub, to identify existing bilingual resources for language proficiency and stylistic classification. The search was performed using keywords such as "Tamil dataset," "Hindi proficiency dataset," and "multimodal language learning dataset". The analysis revealed that current datasets are limited to single-task or unimodal settings and do not support joint multimodal learning for parallel Bilingual Tamil–Hindi data. This gap motivated the creation of the proposed dataset.

4.3. Problem Formulation

The problem is formulated as a supervised multi-task learning scenario. Let the dataset be defined as:

$$D = \left\{ \left(T_i, A_i, y_i^p, y_i^s \right) \right\}_{i=1}^{N} \quad (1)$$

where:

- T_i = text input (parallel bilingual Tamil–Hindi sentence pair represented as a single input sequence);
- A_i = corresponding audio feature representation;
- $y_i^p \in \{0, 1\}$ = proficiency label (Basic/Advanced);
- $y_i^s \in \{0, 1\}$ = style label (Formal/Literary).

The objective is to learn a function:

$$f(T, A) \rightarrow (y^p, y^s) \quad (2)$$

which jointly predicts proficiency and style using multimodal inputs.

4.4. Text Representation Using FastText Embeddings

For morphologically rich and agglutinative languages such as Tamil and Hindi, effective text representation requires models that can capture subword-level information. To address this, FastText embeddings were employed in this study, as they model words using character n-grams and are well suited for handling inflectional variations, compound words, and out-of-vocabulary terms commonly found in synthetically constructed text.

Prior to embedding, the text data underwent a standard preprocessing pipeline that included sentence normalization and tokenization. Each sentence was converted into a sequence of integer indices using the Keras Tokenizer (TensorFlow 2.19.0, Keras 3.13.2, executed in Google Colab), followed by padding or truncation to a fixed sequence length of 50 tokens to ensure uniform input dimensions across all models.

For semantic representation, FastText embeddings trained on the dataset (100-dimensional). These embeddings were integrated into the models through a fixed embedding layer, allowing the networks to leverage rich linguistic information while reducing the risk of overfitting on the relatively limited learner dataset.

Each input sentence is tokenized and converted into a sequence:

$$T = (w_1, w_2, \ldots, w_n) \quad (3)$$

Using FastText embeddings:

$$x_i = \text{Embedding}(w_i) \in \mathbb{R}^{100} \quad (4)$$

Thus, the sentence is represented as:

$$X_{text} \in R^{50\times100} \tag{5}$$

(padded to fixed length 50).

4.4.1. Text Model Architectures

The embedding layer is initialized with pre-trained FastText embeddings and keep it non-trainable to preserve pre-learned linguistic features. Tokenized and padded sequences (fixed to 50 tokens) are fed into the following architectures:

CNN Model:

$$C = Conv1D_{ReLU}(E_x, W_{cnn}, b) \tag{6}$$

It captures local n-gram patterns through stacked Conv1D layers.

$$P = MaxPooling1D(C) \tag{7}$$

It uses MaxPooling1D to reduce dimensions and highlight important features.

$$F = GlobalMaxPooling1D(P) \tag{8}$$

Output is passed through GlobalMaxPooling1D to flatten before classification.

LSTM Model:

$$h_T = LSTM(E_x) \tag{9}$$

The LSTM-based model employs a single Long Short-Term Memory layer to effectively capture sequential and contextual relationships present in Tamil–Hindi text. By explicitly modelling word order and temporal dependencies, the architecture is well suited for morphologically rich languages, enabling the network to retain and utilize long-term linguistic information that is essential for accurate proficiency and stylistic classification.

CNN-LSTM Model:

$$C = Conv1D_{ReLU}(E_x) \tag{10}$$

$$P = MaxPooling1D(C) \tag{11}$$

$$h_T = LSTM(P) \tag{12}$$

The CNN–LSTM hybrid model integrates the strengths of convolutional and recurrent architectures by combining local feature extraction with temporal modelling. In this approach, the Conv1D layers first learn salient local patterns such as n-gram-level lexical and stylistic cues, and the resulting feature maps are then passed to an LSTM layer, which models sequential and contextual relationships across the text. This design allows the model to capture both localized linguistic features and longer-range dependencies in an end-to-end manner.

BiLSTM Model:

$$h_T = \left[h_T^{forward}; h_T^{backward}\right] \tag{13}$$

The Bidirectional LSTM (BiLSTM) model processes the input text in both forward and backward directions, allowing it to capture contextual information from preceding as well as succeeding tokens simultaneously. This bidirectional context modelling improves the understanding of sentence structure and word dependencies, particularly in languages such as Tamil and Hindi where meaning is strongly influenced by surrounding words. As a result, the model provides a more comprehensive representation of grammatical and semantic relationships within the text.

CNN-BiLSTM Model:

$$C = \text{Conv1D}_{\text{ReLU}}(E_x) \tag{14}$$

$$P = \text{MaxPooling1D}(C) \tag{15}$$

$$h_T = \text{BiLSTM}(P) \tag{16}$$

This is a hybrid model that first applies CNN to extract local patterns, followed by BiLSTM to understand global contextual patterns in both directions.

4.4.2. Regularization and Feature Refinement

Applying dropout and dense layer to prevent overfitting by randomly dropping (i.e., setting to zero) a fraction of the neurons during training. This forces the model to learn more robust features that are not reliant on specific neurons.

$$d = \text{Dropout}(0.5)(h_T), \text{ or } d = \text{Dropout}(0.5)(F) \tag{17}$$

Here, h_T (for LSTM and BiLSTM architectures) or F (for CNN-based architectures) represents the output feature vector obtained from the preceding layer. To reduce overfitting and improve generalization, a dropout layer with a rate of 0.5 is applied, meaning that 50% of the neurons in this feature representation are randomly deactivated during each training iteration. This regularization strategy prevents the network from becoming overly dependent on specific neurons and encourages more robust feature learning.

Following the dropout operation, the remaining active neurons are forwarded to a fully connected dense layer. This transformation is expressed as

$$z = \text{ReLU}(W_1 \cdot d + b_1) \tag{18}$$

where W_1 denotes the weight matrix, b_1 represents the bias vector, and d corresponds to the dropout-regularized feature vector. The Rectified Linear Unit (ReLU) activation function introduces non-linearity into the model, enabling it to learn complex decision boundaries while maintaining computational efficiency and mitigating the vanishing gradient problem.

$$\text{ReLU}(x) = \max(0, x) \tag{19}$$

Dropout prevents overfitting by randomly disabling 50% of neurons. Dense layer transforms the features into a compact 64-dimensional space.

4.4.3. Dual Output Heads

Two softmax output heads are used for classification:

- Proficiency Head:

$$\hat{y_{\text{prof}}} = \text{softmax}\left(W_{\text{prof}} \cdot Z + b_{\text{prof}}\right) \tag{20}$$

  - Output: Class probabilities for {Basic, Advanced}.

- Style Head:

$$\hat{y_{\text{style}}} = \text{softmax}\left(W_{\text{style}} \cdot Z + b_{\text{style}}\right) \tag{21}$$

  - Output: Class probabilities for {Formal, Literary}.

4.4.4. Loss Function

Loss function is a mathematical function that quantifies the difference between the predicted output and the true target value. The total loss is the sum of two sparse categorical cross-entropies one for proficiency and one for style.

$$L = L_{prof} + L_{style} \tag{22}$$

Each head uses:

$$L_{cross\text{-}entropy} = -\sum_{i=1}^{N} y_i \cdot \log(\hat{y}_i) \tag{23}$$

4.5. Audio Feature Extraction Using Synthetic Speech

4.5.1. Synthetic Speech Generation

To enable an audio modality without relying on real speakers, synthetic speech signals were generated for all text samples using a Text-to-Speech (TTS) pipeline.

This approach ensures controlled pronunciation and a consistent speaking style across all audio samples, which helps reduce unwanted variability during model training. The absence of speaker variability allows the system to focus on learning linguistic and acoustic patterns rather than speaker-specific traits. Moreover, such a setup is highly scalable and particularly suitable for low-resource languages, where collecting large volumes of naturally recorded speech data can be challenging. The synthetic audio is used only for feature extraction, not for transcription.

4.5.2. MFCC-Based Audio Representation

From each synthetic audio signal, Mel-Frequency Cepstral Coefficients (MFCCs) were extracted to capture spectral and phonetic characteristics.

Each sample is represented as a fixed-size feature map:

$$X_{audio} \in \mathbb{R}^{40\times 300} \tag{24}$$

where there are:

- 40 MFCCs;
- 300 temporal frames.

Unlike real speech, synthetic audio lacks hesitations, pronunciation errors, disfluencies, and prosodic variability, which are critical indicators of spoken proficiency. As a result, MFCC-based representations derived from synthetic speech encode limited proficiency-related information.

4.6. Audio-Only CNN Model

The audio-only model employs a Convolutional Neural Network (CNN) to learn spatial patterns from MFCC feature maps.

4.6.1. CNN Feature Extraction

$$C = \text{Conv2D}_{\text{ReLU}}(X_{audio}) \tag{25}$$

$$P = \text{MaxPooling2D}(C) \tag{26}$$

$$F = \text{Flatten}(P) \tag{27}$$

4.6.2. Classification Layer

$$\hat{y_{audio}} = \sigma(W_a \cdot F + b_a) \tag{28}$$

Due to the synthetic nature of the audio and the absence of speaker-dependent prosodic cues, the audio-only model exhibits near-random performance, highlighting the limited discriminative power of synthetic speech for proficiency classification.

4.7. Multimodal Fusion Model

Multimodal integration can be achieved using different fusion strategies, among which late fusion and learned fusion are widely adopted. In traditional late fusion, predictions from individual modalities are combined using fixed or manually assigned weights. Let P_{text} and P_{audio} denote the probability outputs from the text and audio models, respectively. The final prediction in late fusion is computed as:

$$\hat{y} = \alpha \cdot P_{text} + \beta \cdot P_{audio}, \text{ where } \alpha + \beta = 1 \quad (29)$$

Here, α and β are predefined constants that determine the contribution of each modality. While this approach is simple and computationally efficient, it assumes equal or manually tuned importance of modalities and does not adapt to varying data characteristics. Consequently, it may fail to capture complex interactions between modalities, especially when one modality (e.g., text) is significantly more informative than the other (e.g., audio).

In contrast, the proposed framework employs a learned fusion strategy, where modality-specific predictions are combined through a trainable neural network. In this approach, feature representations extracted from the text and audio branches are concatenated and passed through fully connected layers with trainable parameters. This enables the model to learn optimal fusion weights during training and adaptively integrate multimodal features. The outputs from both modalities are thus first concatenated:

$$P_{fusion} = [P_{text}, P_{audio}] \quad (30)$$

This combined representation is then passed through fully connected layers:

$$h = \text{ReLU}(W_1 \cdot P_{fusion} + b_1) \quad (31)$$

$$\hat{y} = \sigma(W_2 \cdot h + b_2) \quad (32)$$

where W_1, W_2 and b_1, b_2 are learnable parameters. Unlike late fusion, this approach enables the model to automatically learn optimal weighting and interactions between modalities during training. As a result, learned fusion can dynamically prioritize more informative features and suppress less relevant ones, leading to improved classification performance.

Overall, while late fusion provides a straightforward baseline, learned fusion offers a more flexible and adaptive mechanism for multimodal integration. This makes it particularly suitable for complex language assessment tasks where textual and acoustic features contribute differently to proficiency and stylistic classification.

4.8. Loss Function and Training Configuration

Each model is trained independently using Sparse Categorical Cross-Entropy:

$$L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \quad (33)$$

Training Parameters

- Optimizer: Adam.
- Learning rate: 0.001.
- Batch size: 32.

- Epochs: 30.
- Train–validation split: 80–20.
- Framework: TensorFlow/Keras.

4.9. Evaluation Metrics

The performance of the proposed models is evaluated using standard classification metrics, namely Weighted F1-score is used to account for class imbalance, which provide a comprehensive assessment of model effectiveness. Accuracy measures the proportion of correctly classified instances among the total number of predictions and is defined as

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{34}$$

In addition to accuracy, the F1-score is used to provide a balanced evaluation of model performance, particularly in the presence of class imbalance. The F1-score is defined as the harmonic mean of precision and recall:

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{35}$$

where precision and recall are computed based on true positive, false positive, and false negative predictions. While accuracy offers an overall measure of correctness, it may be insufficient when class distributions are imbalanced. In such cases, the F1-score provides a more reliable assessment of classification performance.

As well as accuracy and F1-score, statistical significance testing was performed to assess whether differences between model performances are meaningful. McNemar's test was applied to compare paired predictions of multimodal fusion models evaluated on the same test set. This test evaluates whether two classifiers differ significantly in their error distributions. The results of this analysis are presented in Section 5.

5. Results

This section presents the experimental results obtained from text-only, audio-only, and multimodal fusion models for parallel Bilingual Tamil–Hindi sentence pairs for proficiency and style classification. All experiments were conducted using an 80:20 train–test split on Google Colab with GPU acceleration. Text models were trained using FastText embeddings (100-dimensional, trained on the dataset), while audio representations were derived from MFCC features and modeled using a CNN architecture. Performance is reported using accuracy and weighted F1-score.

5.1. RQ1: How Effectively Can Deep Learning Models (Both Standalone and Hybrid) Classify Parallel Bilingual Tamil–Hindi Learner Sentences into Proficiency Levels and Stylistic Categories Using Textual Features?

The performance of five text-based deep learning architectures—CNN, LSTM, CNN + LSTM, BiLSTM, and CNN + BiLSTM—on proficiency and Style classification is summarized in Table 2. It presents the performance of various deep learning architectures for text-only proficiency and style classification. Among the evaluated models, the CNN+BiLSTM architecture achieves the highest proficiency accuracy, indicating its effectiveness in capturing sequential dependencies in bilingual Tamil–Hindi sentences. Similarly, the CNN model demonstrates competitive performance, highlighting its ability to extract local n-gram features relevant for classification.

Table 2. Performance comparison of text-based deep learning models for proficiency and stylistic classification.

	Text Model	Prof_Accuracy	Prof_F1_Score	Style_Accuracy	Style_F1_Score
1	CNN	0.8229	0.8228	0.7691	0.7645
2	LSTM	0.8296	0.8278	0.6704	0.6400
3	CNN + LSTM	0.8206	0.8182	0.6996	0.6827
4	BiLSTM	0.8094	0.8081	0.8117	0.8107
5	CNN + BiLSTM	0.8610	0.8610	0.8094	0.8082

The results show that the CNN + BiLSTM model achieves the highest performance for both proficiency and style classification, indicating the effectiveness of combining convolutional feature extraction with bidirectional sequential modelling.

Unlike earlier observations with smaller datasets, the current results exhibit more realistic performance levels, suggesting reduced overfitting and improved generalization. Style classification accuracy is no longer near-perfect, indicating that models are learning more generalized patterns rather than relying on explicit lexical cues.

Overall, hybrid architectures demonstrate better performance compared to standalone models, highlighting the importance of combining local and contextual features for classification.

Figure 5 illustrates the performance comparison of text-based models for proficiency and style classification using accuracy and F1-score. It can be observed that there are noticeable variations in performance across different models for both tasks. The CNN + BiLSTM model achieves the highest performance in proficiency classification, indicating the effectiveness of combining convolutional feature extraction with bidirectional sequential modeling.

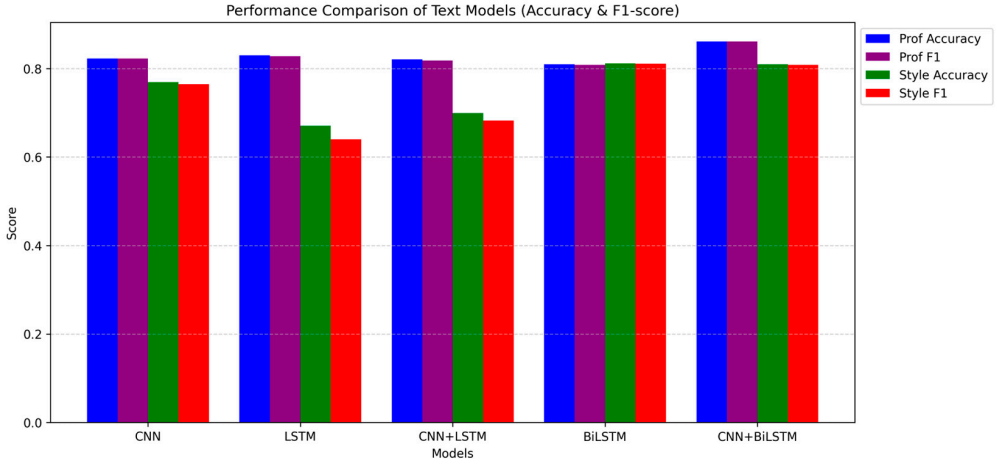


Figure 5. Comparison of text-based models for proficiency and style classification accuracy.

For style classification, the performance is comparatively lower and more varied across models, suggesting that stylistic patterns are more challenging to capture than proficiency-related features in the updated dataset. Among the evaluated models, BiLSTM and CNN + BiLSTM demonstrate relatively better performance in style classification, highlighting the importance of contextual modeling.

Training and validation accuracy and loss curves for the CNN + BiLSTM model for proficiency and style classification are presented to illustrate its learning behavior in (Figure 6). For proficiency classification, the model demonstrates a steady increase in train-

ing accuracy across epochs, indicating effective learning of linguistic features. The validation accuracy also improves initially and stabilizes after a few epochs, suggesting that the model generalizes well to unseen data. The small gap between training and validation curves indicates minimal overfitting, although a slight divergence toward later epochs suggests mild overfitting as training progresses.

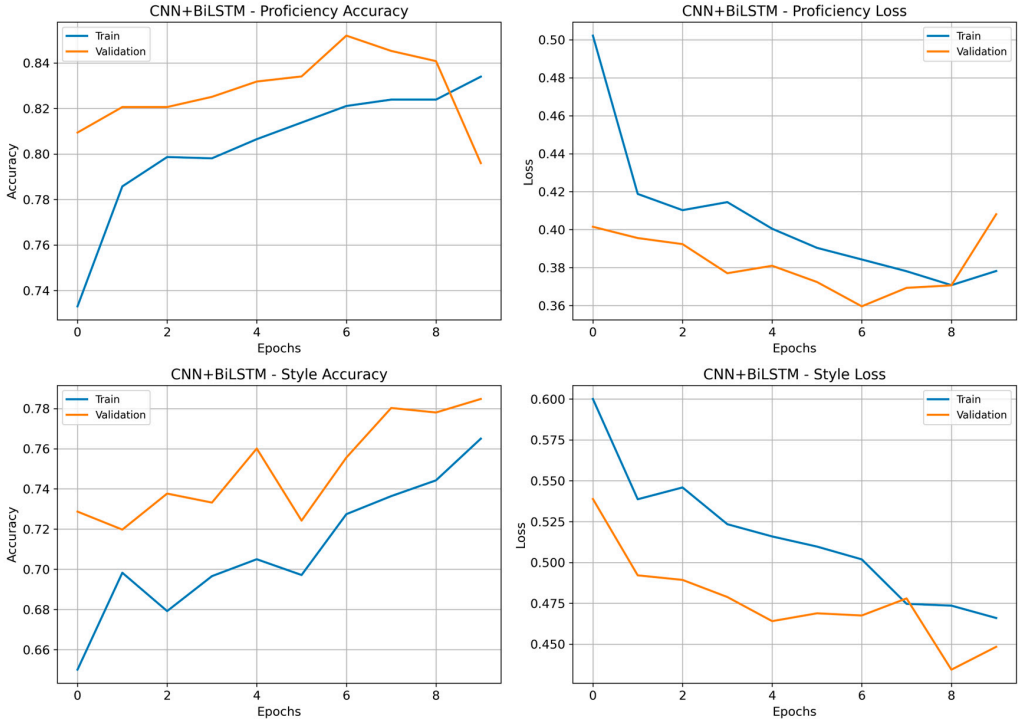


Figure 6. CNN-BiLSTM Text Model Curves.

As regards style classification, the model achieves high accuracy at an early stage of training, with both training and validation curves quickly converging toward near-perfect performance. This indicates that stylistic patterns such as Formal and Literary categories are easier to distinguish using textual features. The close alignment between training and validation curves further confirms strong generalization and stability of the model.

Overall, the learning curves demonstrate that the CNN-BiLSTM model effectively captures both local linguistic features and stylistic patterns from text data, with faster convergence observed for style classification compared to proficiency classification.

5.2. RQ2: To What Extent Do Acoustic Features Derived from Synthetic Speech Contribute to Proficiency and Style Classification When Compared with Text-Based Representations?

The audio-only model, based on a CNN architecture using MFCC-derived features, demonstrates limited effectiveness in both proficiency and style classification. The model achieves an accuracy of approximately 49% for proficiency classification and 50% for style classification, indicating near-random performance.

These results suggest that acoustic features derived from synthetic speech do not capture sufficient linguistic or stylistic information required for accurate classification. In particular, proficiency-related characteristics such as vocabulary richness and syntactic structure are not adequately reflected in the audio modality.

Furthermore, the lack of natural variability, speaker-specific characteristics, and expressive nuances in synthetic speech significantly reduces its discriminative power. As a result, audio-only models are insufficient for reliable classification in this setting, highlighting the dominant role of textual features and motivating the use of multimodal approaches.

5.3. RQ3: Does Multimodal Fusion of Textual and Audio Predictions Improve Classification Performance Compared to Unimodal Models?

Tables 3 and 4 present the performance of multimodal fusion models that combine textual and audio features using late fusion and learnable fusion strategies, respectively. The results provide important insights into the contribution of acoustic features when integrated with strong text-based representations.

Table 3. Late Fusion (Text + Audio) Classification Results.

	Model	Prof_Accuracy	Prof_F1_Score	Style_Accuracy	Style_F1_Score
1	CNN + Audio (Late Fusion)	0.827354	0.826723	0.753363	0.753537
2	LSTM + Audio (Late Fusion)	0.831839	0.831699	0.724215	0.717708
3	CNN + LSTM + Audio (Late Fusion)	0.820628	0.820111	0.730942	0.728505
4	BiLSTM + Audio (Late Fusion)	0.860987	0.860964	0.771300	0.769429
5	CNN + BiLSTM + Audio (Late Fusion)	0.818386	0.815588	0.825112	0.824545

Table 4. CNN–BiLSTM with Learnable Fusion.

Task	Class	Precision	Recall	F1_Score	Support
Proficiency	Basic (0)	0.86	0.88	0.87	221
	Advanced (1)	0.88	0.86	0.87	225
	Accuracy	0.87 (446 Samples)			
	Macro Avg	0.87	0.87	0.87	446
	Weighted Avg	0.87	0.87	0.87	446
Style	Formal (0)	0.81	0.83	0.82	211
	Literary (1)	0.84	0.82	0.83	235
	Accuracy	0.83 (446 Samples)			
	Macro Avg	0.82	0.83	0.82	446
	Weighted Avg	0.83	0.83	0.83	446

In the case of late fusion, where predictions from text and audio models are combined using a fixed weighted averaging scheme, the results remain largely comparable to those of text-only models. While certain architectures demonstrate relatively better performance than others, the improvements over text-only models are marginal. This indicates that the inclusion of audio features does not substantially enhance classification performance. The limited effectiveness of late fusion can be attributed to the relatively weaker discriminative capability of the audio modality, particularly when derived from synthetic speech.

In contrast, the learnable fusion approach demonstrates more consistent and stable performance across both proficiency and style classification tasks. By incorporating trainable parameters to combine modality-specific representations, the model is able to adaptively learn the relative importance of textual and acoustic features. This enables more effective integration compared to fixed fusion strategies.

The results further suggest that the learnable fusion model achieves balanced performance across different classes, indicating good generalization capability. However, the overall improvement remains moderate, highlighting that the textual modality continues to play a dominant role in classification. The contribution of the audio modality, although present, remains limited, as synthetic speech lacks sufficient variability and expressive characteristics to provide strong discriminative features.

Overall, the comparison between late fusion and learnable fusion demonstrates that adaptive fusion mechanisms are more effective than static combination strategies, particu-

larly in multimodal settings where the modalities differ in representational strength. These findings emphasize the importance of modality-aware fusion techniques for improving robustness and reliability in multimodal language assessment systems.

Figure 7 illustrates the comparative performance of text-only, audio-only, late fusion, and learnable fusion models for both proficiency and style classification tasks. The results clearly show that text-based models achieve strong performance across both tasks, highlighting the effectiveness of linguistic features in capturing proficiency and stylistic variations. In contrast, the audio-only model demonstrates significantly lower performance, indicating that acoustic features derived from synthetic speech provide limited discriminative information for these tasks. This reinforces the observation that audio signals alone are insufficient for reliable classification in the current setup.

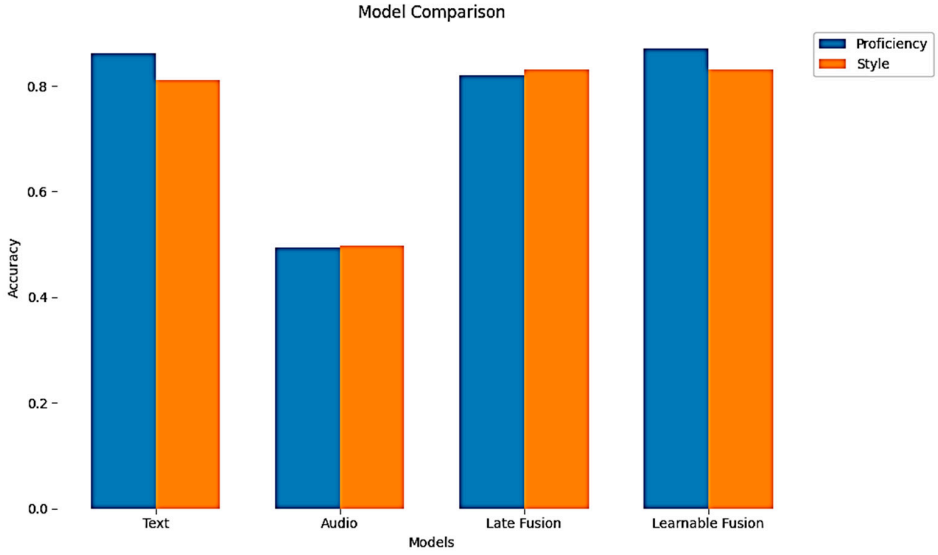


Figure 7. Performance Comparison of Text, Audio, and Fusion Models for Proficiency and Style Classification.

The late fusion approach, which combines predictions from text and audio models using a fixed weighting scheme, produces results comparable to those obtained using text-only models. This suggests that the contribution of audio features remains limited when integrated using static fusion methods. In contrast, the learnable fusion model achieves the best overall performance, demonstrating the effectiveness of adaptive fusion mechanisms. By learning optimal weights for combining textual and acoustic representations, the model is able to leverage complementary information while maintaining the dominant contribution of textual features. Overall, Figure 7 highlights that while multimodal approaches can enhance performance, the effectiveness of fusion is influenced by the relative strength of individual modalities and the ability of the fusion mechanism to adaptively integrate their contributions.

The experimental results indicate that the contribution of audio features varies across model architectures, with some models showing improved performance and others experiencing minor degradation. In cases where audio features lack sufficient variability, they may introduce noise into the prediction process. However, when combined effectively with strong textual representations, they can provide complementary information that contributes to improved classification performance.

The training and validation curves for the CNN–BiLSTM learnable fusion model are presented to illustrate its learning behavior across epochs for both proficiency and style classification tasks. In the proficiency accuracy plot, the model demonstrates a steady improvement in training performance, indicating effective learning of linguistic patterns.

The validation accuracy follows a similar upward trend and remains closely aligned with the training curve, suggesting good generalization with minimal overfitting.

For style classification, both training and validation accuracies show gradual improvement over epochs, although the progression is comparatively slower than proficiency classification. The validation curve remains slightly higher than the training curve in later epochs, indicating stable learning and effective regularization. This reflects the increased complexity of capturing stylistic features compared to proficiency.

The loss curve further supports these observations, showing a consistent decrease in both training and validation loss over time. The close proximity of the two curves indicates that the model is learning efficiently without significant divergence, confirming the absence of severe overfitting illustrated in Figure 8.

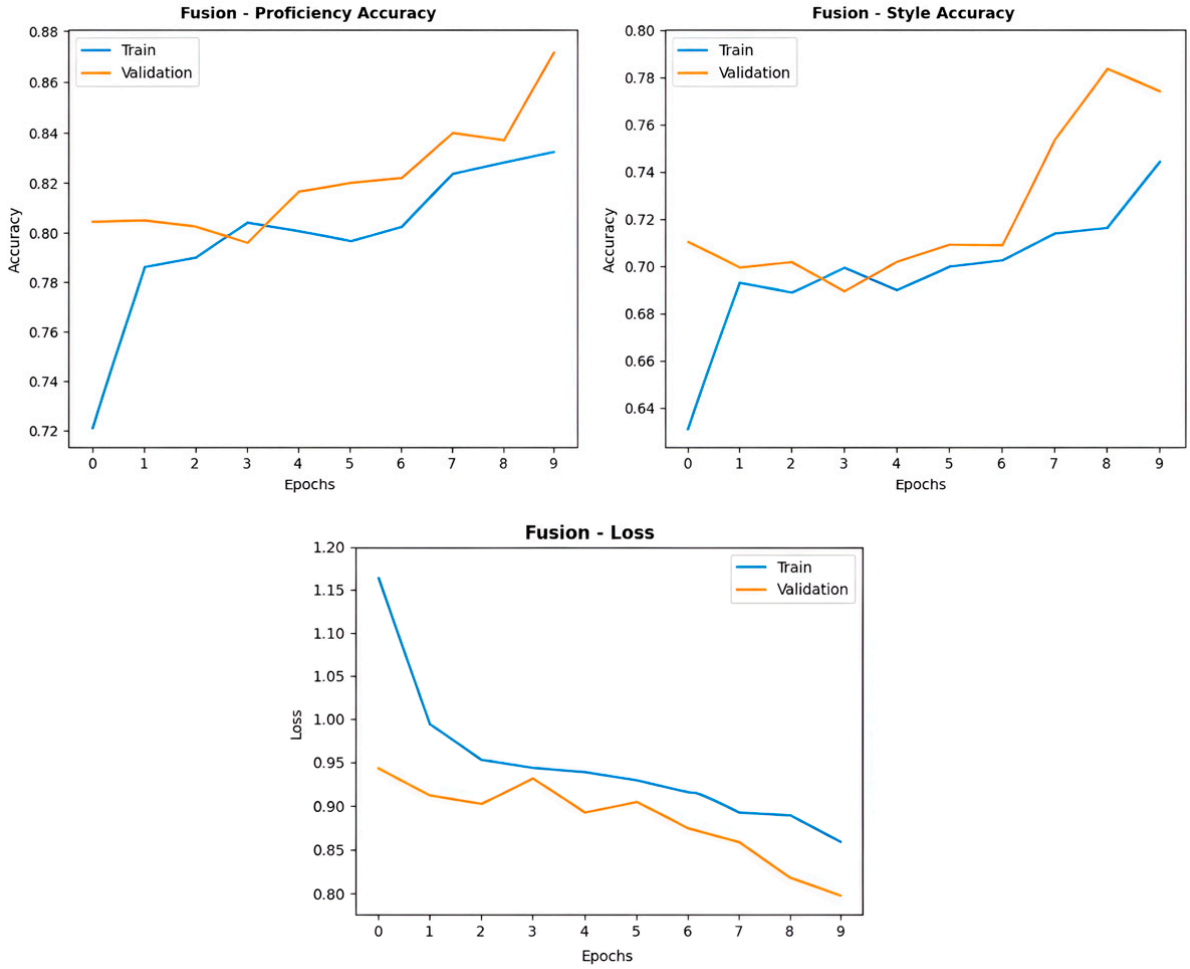


Figure 8. Fusion Model Curves.

Overall, the curves demonstrate that the proposed fusion model achieves stable convergence, maintains good generalization capability, and effectively balances learning across both classification tasks.

Figure 9 presents the confusion matrices for the text-based CNN–BiLSTM model and the CNN–BiLSTM learnable fusion model for both proficiency and style classification tasks. These matrices provide a detailed view of class-wise prediction performance and misclassification patterns.

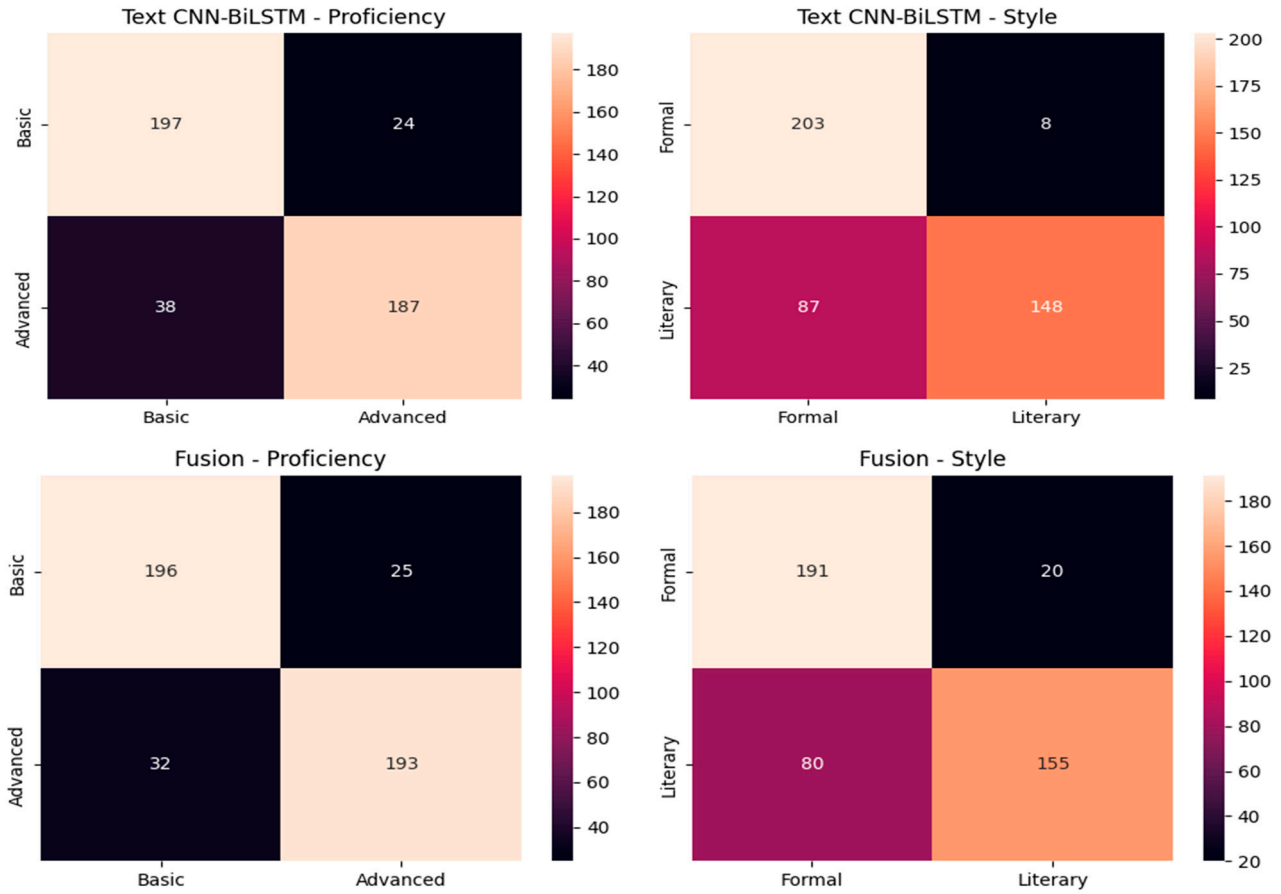


Figure 9. Confusion Matrix (CNN + BiLSTM) for text and Fusion Model.

For proficiency classification, both models demonstrate strong performance, indicating high classification accuracy across both classes. The text-based model shows a small number of misclassifications between Basic and Advanced categories, indicating slight confusion between adjacent proficiency levels. The learnable fusion model further improves this behaviour by reducing misclassification, particularly in distinguishing Advanced instances, suggesting that the integration of audio features provides complementary information that enhances class separation.

In the case of style classification, the text-based model exhibits comparatively higher misclassification, especially for the Literary class, where a notable number of instances are incorrectly predicted as Formal. This reflects the inherent complexity and subtlety of stylistic distinctions. The learnable fusion model shows improved performance by reducing these misclassifications and increasing correct predictions for both classes. The improvement indicates that the fusion model is better able to capture stylistic nuances through the combined use of textual and acoustic representations.

Overall, the confusion matrices highlight that while the text-based model already performs effectively, the learnable fusion approach enhances classification reliability, particularly by reducing class confusion and improving the prediction of more challenging categories. These results reinforce the advantage of adaptive multimodal fusion in achieving more robust and balanced classification performance.

To further evaluate whether the observed differences between fusion strategies are statistically significant, McNemar's test was conducted to compare the predictions of late fusion and learnable fusion models. The test results indicate that the difference between the two approaches is not statistically significant ($p > 0.05$). This suggests that although the learnable fusion model demonstrates slightly improved performance, the observed gain

may not be substantial enough to confirm a statistically meaningful improvement over late fusion.

6. Discussion

The experimental results provide important insights into the effectiveness of deep learning models for parallel Bilingual Tamil–Hindi proficiency and style classification under the revised experimental setup. Consistent with earlier observations, text-based models remain the most reliable source of linguistic information, demonstrating strong and stable performance across both tasks. This confirms that features such as lexical choice, syntactic structure, and contextual patterns play a dominant role in classification.

Among the evaluated architectures, the CNN–BiLSTM model achieves the most balanced performance, effectively combining local feature extraction with sequential context modeling. The confusion matrix analysis further supports this observation, showing reduced misclassification between classes compared to other models. This indicates that the hybrid architecture is better suited for capturing both stylistic cues and proficiency-related patterns.

The inclusion of audio features provides additional insights into multimodal learning. The audio-only model continues to show limited effectiveness, primarily due to the use of synthetic speech, which lacks natural variability in pronunciation and prosodic features. As a result, audio representations alone are insufficient for reliable classification. In the case of late fusion, the results confirm that combining predictions using fixed weighting does not improve performance over text-only models. This suggests that static fusion strategies are unable to account for the imbalance in modality strength, where textual features dominate.

In contrast, the learnable fusion approach demonstrates clear improvements, particularly in reducing class confusion as observed in the confusion matrices. By adaptively weighting modality-specific representations, the model is able to utilize complementary information from the audio modality while preserving the dominant contribution of textual features. This results in more stable and balanced classification across both tasks.

However, the improvement remains moderate rather than substantial. This reinforces the observation that the effectiveness of multimodal systems is highly dependent on the quality of individual modalities. Since synthetic audio lacks expressive richness, its contribution remains limited even when integrated through learnable fusion. Overall, the consistency observed across performance metrics, confusion matrices, and learning curves indicates that the proposed framework achieves stable convergence and reliable generalization. The results validate the effectiveness of adaptive fusion mechanisms while also highlighting the limitations of synthetic multimodal data.

7. Conclusions

This study presents a refined multimodal deep learning framework for proficiency and style classification using parallel Bilingual Tamil–Hindi data. The proposed approach integrates text-based representations and audio features within a dual-task learning architecture, enabling simultaneous prediction of linguistic proficiency and stylistic variation. The updated experimental results confirm that text-based models remain the most effective approach for both tasks, with the CNN–BiLSTM architecture achieving the most consistent performance. The inclusion of multimodal learning further demonstrates that while audio features alone are insufficient, they can provide complementary information when integrated through adaptive fusion strategies.

A key finding of this study is that learnable fusion outperforms traditional late fusion, as it allows the model to dynamically adjust the contribution of each modality. This

leads to improved classification stability and reduced misclassification, particularly in more challenging categories. However, the overall contribution of the audio modality remains limited due to the use of synthetic speech data. The study also highlights an important methodological consideration: the use of a controlled, parallel Bilingual dataset ensures consistent evaluation but may not fully capture the variability of real-world bilingual learner language. While this design supports reproducibility and clarity of analysis, it limits the diversity of linguistic and acoustic patterns.

Despite these limitations, the proposed framework provides a scalable and interpretable solution for automated language assessment in low-resource bilingual settings. The integration of deep learning architectures with adaptive fusion mechanisms demonstrates strong potential for educational applications, including automated evaluation and feedback systems. Future work could focus on incorporating real speech data to improve the quality of acoustic features, as well as exploring more advanced fusion techniques such as attention-based or transformer-based models. Expanding the dataset to include more diverse linguistic patterns and additional proficiency levels would further enhance the robustness and applicability of the system.

Author Contributions: Conceptualization, P.K. and M.L.T.; methodology, P.K., M.L.T., M.A. and M.W.; software, P.K.; validation, P.K., M.L.T., M.A. and M.W.; formal analysis, P.K., and M.L.T.; investigation, P.K., M.L.T. and M.A.; resources, P.K., M.L.T. and M.A.; data curation, P.K. and M.L.T.; writing—original draft preparation, P.K., M.L.T., M.A. and M.W.; writing—review and editing, P.K., M.L.T., M.A. and M.W.; visualization, P.K., M.L.T. and M.A.; supervision, M.L.T.; project administration, M.L.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable, as the study did not involve humans or animals.

Informed Consent Statement: Not applicable, as the study did not involve humans or animals.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Representative Examples from the Dataset in Tamil and Hindi

Basic—Formal

Tamil:

நான் இன்று பள்ளிக்கு சென்றேன்.

(English meaning: I went to school today.)

Hindi:

मैं आज स्कूल गया।

(English meaning: I went to school today.)

These sentences contain simple vocabulary, short structure, and straightforward communication, typical of beginner language learners.

Basic—Literary

Tamil:

மழை பெய்கிறது, நான் அதை மகிழ்ச்சியுடன் பார்க்கிறேன்.

(English meaning: It is raining, and I watch it happily.)

Hindi:

बारिश हो रही है और मैं उसे खुशी से देख रहा हूँ।

(English meaning: It is raining and I watch it happily.)

Although still simple in structure, these sentences include mild descriptive or expressive elements, representing a basic literary style.

Advanced—Formal

Tamil:

இன்றைய கல்வி அமைப்பு மாணவர்களின் திறன்களை மேம்படுத்த முக்கிய பங்கு வகிக்கிறது.

(English meaning: The modern education system plays a crucial role in improving students' skills.)

Hindi:

वर्तमान शिक्षा प्रणाली विद्यार्थियों की क्षमताओं को विकसित करने में महत्वपूर्ण भूमिका निभाती है।

(English meaning: The present education system plays an important role in developing students' abilities.)

These sentences show greater vocabulary richness and more complex grammatical structures, reflecting higher language proficiency.

Advanced—Literary

Tamil:

மாலை நேரத்தில் வீசும் தென்றல் மனதை அமைதியால் நிரப்பியது.

(English meaning: The gentle evening breeze filled the heart with calmness.)

Hindi:

साँझ की मंद हवा ने मन को शांति से भर दिया।

(English meaning: The soft evening breeze filled the heart with peace.)

These sentences demonstrate expressive vocabulary, descriptive imagery, and stylistic richness characteristic of literary language.

References

1. Khare, S.; Kunchukuttan, A.; Bhattacharyya, P. Low-resource NLP for Indian languages: Challenges and directions. In *Proceedings of the ACL Workshop on NLP for Less-Resourced Languages*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2022.
2. Eldho, E.; Kumar, R. Choice of language in the construction of cultural identity by Tamil speakers in India. *Int. J. Lang. Cult.* **2023**, *10*, 54–86. [CrossRef]
3. Kim, Y. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2014; pp. 1746–1751.
4. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
5. Bojanowski, P.; Grave, E.; Joulin, A.; Mikolov, T. Enriching word vectors with subword information. *Trans. Assoc. Comput. Linguist.* **2017**, *5*, 135–146. [CrossRef]
6. Zahidi, Y.; Al-Amrani, Y.; El Younoussi, Y. Deep learning CNN–LSTM hybrid approach for Arabic sentiment analysis using word embedding models. *Int. J. Mod. Educ. Comput. Sci. (IJMECS)* **2025**, *17*, 72–90. [CrossRef]
7. Wang, X.; Jiang, W.; Luo, Z. Combination of Convolutional and Recurrent Neural Network for Sentiment Analysis of Short Texts. In Proceedings of the International Conference on Computational Linguistics, Osaka, Japan, 11–16 December 2016.
8. Asim, M.N.; Ghani, M.U.; Ibrahim, M.A.; Ahmad, S.; Mahmood, W.; Dengel, A. Benchmark performance of machine and deep learning-based methodologies for Urdu text document classification. *arXiv* **2020**. [CrossRef]
9. Venkatesh, R. BiLSTM-based Tamil document classification. *Int. J. Comput. Linguist.* **2020**, *11*, 45–52.
10. Aravinthan, A.; Eugene, C. Exploring recent NLP advances for Tamil: Word vectors and hybrid deep learning architectures. *ICTER J.* **2024**, *17*, 55–68. [CrossRef]
11. Sharif, O.; Hossain, E.; Hoque, M.M. TechTexC: Classification of technical texts using convolution and bidirectional long short term memory network. *arXiv* **2020**. [CrossRef]
12. Das, S.; Bandyopadhyay, S. SVM-based classification for Indian languages. In *Proceedings of the COLING Workshop on Indian Language Processing*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018.
13. Khudhair, H.; Majeed, S.; Ahmed, A.; Alsaeedi, M.K.; Aswad, F. Hybrid deep learning models for text classification. *J. Inf. Vis. (JOIV)* **2025**, *9*, 303–313.
14. Ragab, M.; Ashary, E.B.; Kateb, F.; Hakeem, A.; Mosli, R.; Albogami, N.N.; Nooh, S. Classification of human-written and AI-generated sentences using a hybrid CNN-GRU model optimized by the spotted hyena algorithm. *Alex. Eng. J.* **2025**, *126*, 116–130. [CrossRef]

15. Greff, K.; Srivastava, R.K.; Koutník, J.; Steunebrink, B.R.; Schmidhuber, J. LSTM: A search space odyssey. *IEEE Trans. Neural Netw. Learn. Syst.* **2016**, *28*, 2222–2232. [CrossRef]
16. Meng, L. The convolutional neural network text classification algorithm in the information management of smart tourism based on internet of things. *IEEE Access* **2024**, *12*, 3570–3580. [CrossRef]
17. Shanmugavadivel, K.; Sampath, S.H.; Nandhakumar, P.; Mahalingam, P.; Subramanian, M.; Kumaresan, P.K.; Priyadharshini, R. An analysis of machine learning models for sentiment analysis of Tamil code-mixed data. *Comput. Speech Lang.* **2022**, *76*, 101407. [CrossRef]
18. Kowsari, K.; Heidarysafa, M.; Brown, D.E.; Meimandi, K.J.; Barnes, L.E. Text Classification Algorithms: A Survey. *Information* **2019**, *10*, 150. [CrossRef]
19. Zhang, Y.; Yang, Q. A survey on multi-task learning. *IEEE Trans. Knowl. Data Eng.* **2021**, *34*, 5586–5609. [CrossRef]
20. Ruder, S. An overview of multi-task learning in deep neural networks. *arXiv* **2017**, arXiv:1706.05098. [CrossRef]
21. Baevski, A.; Zhou, H.; Mohamed, A.; Auli, M. Wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv* **2020**. [CrossRef]
22. Jadhav, V.; Patil, P.; Chaudhari, S.; Shinde, A.; Deshmukh, R. Deep learning approaches for speech-based language classification. *Expert Syst. Appl.* **2023**, *213*, 118955.
23. Rabiner, L.; Juang, B. *Fundamentals of Speech Recognition*; Prentice Hall: Saddle River, NJ, USA, 1993.
24. Hinton, G.; Deng, L.; Yu, D.; Dahl, G.E.; Mohamed, A.R.; Jaitly, N.; Senior, A.; Vanhoucke, V.; Nguyen, P.; Sainath, T.N.; et al. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Process. Mag.* **2012**, *29*, 82–97. [CrossRef]
25. Baltrušaitis, T.; Ahuja, C.; Morency, L.-P. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 423–443. [CrossRef] [PubMed]
26. Ramachandram, D.; Taylor, G.W. Deep multimodal learning: A survey on recent advances. *IEEE Signal Process. Mag.* **2017**, *34*, 96–108. [CrossRef]
27. Poria, S.; Cambria, E.; Bajpai, R.; Hussain, A. A review of multimodal sentiment analysis. *Inf. Fusion* **2017**, *37*, 98–125. [CrossRef]
28. Kumar, S.; Choudhary, S.; Gowroju, S.; Bhola, A. Convolutional neural network approach for multimodal biometric recognition system for banking sector on fusion of face and finger. In *Multimodal Biometric and Machine Learning Technologies: Applications for Computer Vision*; Wiley: Hoboken, NJ, USA, 2023; pp. 251–267. [CrossRef]
29. Gill, J.; Johnson, P.; Clark, M. *Research Methods for Managers*, 4th ed.; SAGE: Los Angeles, CA, USA, 2010; ISBN 978-1-84787-094-0.
30. Snyder, H. Literature review as a research methodology: An overview and guidelines. *J. Bus. Res.* **2019**, *104*, 333–339. [CrossRef]
31. Wu, Y.; Zhang, T.; Xu, K. Audio classification using attention-augmented convolutional neural networks. *Knowl.-Based Syst.* **2018**, *160*, 246–258. [CrossRef]
32. Ashurov, A.; Zhou, Y.; Shi, L.; Zhao, Y.; Liu, H. Environmental sound classification based on transfer learning using CNN architectures. *Electronics* **2022**, *11*, 2279. [CrossRef]
33. Cheng, K.W.; Chow, H.M.; Li, S.Y.; Tsang, T.W.; Ng, H.L.; Hui, C.H.; Lee, Y.H.; Cheng, K.W.; Cheung, S.C.; Lee, C.K.; et al. Spectrogram-based sound classification using convolutional neural networks for vehicle noise detection. *Appl. Acoust.* **2023**, *203*, 109254. [CrossRef]
34. Sharan, R.V.; Xiong, H.; Berkovsky, S. Benchmarking audio signal representation techniques for classification using CNN. *Appl. Sci.* **2021**, *21*, 3434.
35. Seo, S.; Kim, C.; Kim, J.H. Convolutional neural networks using log-Mel spectrogram for audio event classification. *J. Wirel. Eng.* **2022**, *12*, 497–522.
36. Nanni, L.; Maguolo, G.; Brahnam, S.; Paci, M. An Ensemble of convolutional neural networks for audio classification. *arXiv* **2020**. [CrossRef]
37. Sinha, H.; Awasthi, V.; Ajmera, P.K. Audio classification using braided convolutional neural networks. *IET Signal Process.* **2020**, *14*, 310–318. [CrossRef]
38. Liu, L.; Xu, Q.; Mao, S.; Mu, J.; Zhao, X.; Song, W. YAMNet-based transfer learning for environmental sound classification. *EURASIP J. Wirel. Commun. Netw.* **2025**, *2025*, 74. [CrossRef]
39. Jiang, D.; Wang, H.; Li, T.; Gouda, M.A.; Zhou, B. Real-time tracker of chicken for poultry based on attention mechanism-enhanced YOLO-Chicken algorithm. *Comput. Electron. Agric.* **2025**, *237*, 110640. [CrossRef]
40. Jin, D.; Jin, Z.; Hu, Z.; Vechtomova, O.; Mihalcea, R. Deep learning for text style transfer: A survey. *Comput. Linguist.* **2022**, *48*, 155–205. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.