SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V13I9P110

Volume 13 Issue 9, 160-180, September 2026

Original Article

A Context-Aware, Feature-Enriched Bidirectional Long Short-Term Memory Framework for Automatic Document Readability Assessment

Rajesh Kandakatla[1], V. Dhilip kumar[2]

[1]*Computer Science and Engineering, Vel Tech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology, Chennai, Tamilnadu, India.*
[2]*Vel Tech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology, Chennai, Tamilnadu, India.*

[1]*Corresponding Author: rajesh.kandakatla@gmail.com*

Received: 30 July 2026 Revised: 16 September 2026 Accepted: 18 September 2026 Published: 29 September 2026

Abstract *- Automatic readability assessment is important in the domain of educational content selection, language learning, digital accessibility, and personalized reading assistance. The traditional readability formulas quickly show a high dependence on bare-bones factors, and as such, they can not adequately assess the context of meaning, syntactic config, semantic entropy, discourse cohesion, and arranged structure types for degrees of ease. This work presents a framework for automatic document readability assessment which takes into consideration the context and other related features. They mixed several elements in their model, including RoBERTa-based contextual embeddings, a Bidirectional Long Short-Term Memory (Bi-LSTM) encoder, an attention mechanism, multidimensional linguistic features, gated feature fusion, and an ordinal-aware learning objective. These lexical, syntactic, semantic, and discourse cues are fused using an attention-weighted contextual representation to obtain a document-level feature vector. In the two-stage transfer-learning strategy, first, we pre-train our encoder on the CEFR-Based Sentence Profile corpus and then fine-tune it with the OneStopEnglish document-level corpus. The final objective involves the use of weighted categorical cross-entropy in conjunction with an ordinal-distance penalty to take into account class imbalance, as well as the seriousness of the error between more distant readability classes. Evaluation metrics: The evaluation metrics used include accuracy, macro-F1, Mean Absolute Error (MAE), Quadratic Weighted Kappa (QWK), adjacent level accuracy, and Spearman rank correlation. The approximation evaluation shows that after using the CEFR-SP pretrained model, our accuracy is 93.19%, macro-F1 is 92.96%, MAE is 0.078, and QWK is 0.956. Ablation studies also show that the model performance increases progressively through incorporating bidirectional encoding, attention, linguistic-feature fusion, auxiliary pretraining, and ordinal supervision. This framework forms a basis for effective context-sensitive and order-aware document readability prediction.*

Keywords *- Automatic readability assessment, BiLSTM, attention mechanism, Linguistic-feature fusion, Ordinal classification, RoBERTa, CEFR-SP, OneStopEnglish.*

1. Introduction

Automatic Readability Assessment (ARA) aims to estimate the readability of a text and enables personalized learning, language learning, digital accessibility, infusion of pedagogical lessons in curriculum development, and content recommendation. Such traditional formulas are based solely on surface-level features, such as the average number of words per sentence or syllables per word, and cannot properly account for semantic, syntactic, contextual, or discourse-level complexity [25, 26]. Both neural and hybrid methods have improved readability modeling by modeling context and linguistic features [1, 4, 9, 15]. However, modeling long context ranges, identifying significant textual fragments, aligning various types of features, and maintaining the ordered nature of readability values remain open problems.

However, with all these advances, there is still a crucial research gap in creating a unified framework for readability assessment that, at the same time, can model long-range bidirectional context, isolate readability-sensitive textual segments, combine heterogeneous linguistic information and contextual information into a single representation, and maintain the order among readability levels. Current methods usually focus on one or more of these aspects separately: recurrent models can cope with the sequentiality, but may lack the capability to capture the most informative part of the text; transformer-based models can give rich context representation, but may not explicitly capture interpretable information; hybrid models may combine neural and handcrafted features, but may not have adaptive feature combination; conventional classification objectives usually assume equally severe adjacent and distant errors. Moreover, the transfer of knowledge of sentence level linguistic difficulty into document level readability assessment is not well studied. Hence, a context-aware, feature-rich framework that simultaneously represents contextual, linguistic and ordinal information with the

ability to transfer knowledge across text granularities is required. This study aims to alleviate this gap by proposing a RoBERTa-based BiLSTM with attention, gated linguistic-feature fusion, CEFR-SP pretraining, and ordinal-aware optimization for document-level readability assessment.

1.1. Background and Motivation

Readability assessment helps provide insight into the extent to which the books align with readers' proficiency and comprehension. CEFR-based and multilingual studies suggest that text difficulty is linked with broadly characterized levels, rather than a general encompassing score [17, 32, 33, 36]. This also facilitates the accessibility of digital and educational content for children, second-language learners and readers with limited domain knowledge [39, 49].

Readability formulas such as Flesch–Kincaid, SMOG, Gunning Fog and Coleman–Liau are simplistic and more transparent but can fail to capture contextual meaning, syntactic ambiguity, conceptual density and discourse organization [25, 26]. Thus, lexical, syntactic, semantic, and discourse information are needed for modern readability assessment. Various features have contributed to predicting text difficulty, such as syntactic structures [7], context embeddings [19, 20], and discourse features [31, 34, 35]. The results of hybrid systems also indicate that explicit linguistic features and neural representations were capturing complementary information [1, 9, 15, 40].

The grades of readability further constitute an ascending grade. Although predicting an adjacent class is less severe than predicting a far-off one, conventional classifiers generally ignore the correlation between classes and treat them as independent.

This has therefore motivated the introduction of ordinal regression, pairwise ranking, soft labels, and distance-sensitive objectives to preserve relationships among difficulty levels [2, 8, 10, 14]. Such limitations motivate the use of a single holistic framework that combines contextual encoding, attention-based mechanisms, linguistic-feature fusion, and ordinal-aware learning.

1.2. Problem Statement

There are four major shortcomings with existing readability-assessment methods. First, unidirectional or shallow models cannot capture bidirectional and long-range dependencies in a long document [21, 34]. Second, since many of these models do not explicitly point to the words or sentences that drive text difficulty, the predictive focus and interpretability drop [1, 12].

Greenspan et al. mention that handcrafted linguistic indicators and neural representations are often employed in a stand-alone manner, which leads to partial modeling of explicit and latent readability information [9, 15, 18, 40]. Fourth, the typical classification objective treats ordered readability levels as nominal categories and does not distinguish between adjacent and distant errors [2, 10, 14].

As such, in this work, we solve the task of document-readability modeling by proposing an approach that extracts bidirectional context, identifies effective text segments, integrates heterogeneous linguistic and neural features, and retains the ordinal relationship of readability scores.

1.3. Research Challenges

The principal research challenges are:

1. Modeling long-range dependencies of context [12, 22];
2. Combining neural embedding with lexical, syntactic, semantic, and discourse features at different scales [1, 6, 18];
3. Discriminating neighboring levels of readability with similar linguistic attributes [36];
4. Respecting the order between the difficulty classes [2, 8, 10, 14];
5. Dealing with the class imbalance problem as well as corpus-related differences [3, 13, 18, 24];
6. Producing explainable predictions by means of attention mechanism and linguistic features [1, 49]; and
7. Transferring readability knowledge to text with other granularities, domains, and even other corpora [3, 17, 24].

The unique aspect of the proposed framework is the combination of four complementary aspects of readability assessment: the contextual and sequential representation learning, the explicit multidimensional linguistic modeling, the adaptive feature integration and the ordinal-aware prediction. The method proposed in this paper integrates the RoBERTa-contextual embedding, BiLSTM-based bidirectional encoding, attention-based document representation, gated fusion of lexical, syntactic and semantic and discourse features, CEFR-SP sentence-level pretraining and the ordinal-distance-aware objective function in one single framework. This design aims to enhance not only the exact readability classification but also the ordinal quality of predictions and/or the translatability of the knowledge of the linguistic difficulty from the sentence level to the document level.

1.4. Main Contributions

The main contributions of this study are:

1. A bidirectional sequential encoder that encodes documents via previous and next contextual information.
2. An attention mechanism that detects words, phrases, and sentences that are sensitive to readability.
3. An integrated lexical, syntactic, semantic, and discourse indicator with contextual neural representations via a feature-fusion module.
4. A distance penalty that is designed for ordered readability classes on top of a weighted cross-entropy objective
5. A transfer learning method wherein the sentence encoder will use the CEFR-SP data while the document level fine tuning and testing will utilize OneStopEnglish.
6. An evaluation framework that considers the following metrics: accuracy, macro-F1, mean absolute error,

adjacent level accuracy, Spearman's rank correlation, and quadratic weighted kappa.
7. Ablation experiments of contextual encoding, attention, feature fusion, ordinal supervision and CEFR-SP pretraining.

2. Related Work

Automatic readability assessment has progressed from formula-based methods to machine-learning, deep-learning, hybrid, and ordinal approaches. A recent trend has been to use contextual embeddings, linguistic features, pretrained language models, and distance-sensitive objectives.

2.1. Traditional Readability Formulas

Conventional measures like Flesch Reading Ease, Flesch–Kincaid Grade Level, Gunning Fog index, SMOG, and Coleman–Liau give a guess as to text reading levels based on statistics of sentence length, word length, number of syllables per word, frequency of polysyllabic words, and such.

They are transparent and computationally efficient, but fail to capture semantic meaning, syntactic complexity, discourse coherence, or reader-specific knowledge [25, 26]. Although they are valuable as baseline measures, they lack sufficient detail for fine-grained readability assessment.

2.2. Machine-Learning Approaches

Machine-learning algorithms employ classifiers like logistic regression, Support Vector Machines, Random Forest, and XGBoost using derived lexical, syntactic, and discourse characteristics. Common features are word count, word frequency, lexical diversity, sentence length, part-of-speech distribution, clause density, and so on [7, 19, 31, 35]. These models can encode complex relations among linguistic variables, but their performance is sensitive to feature quality and may vary across domains or corpora [3, 24]. They also do badly at representing latent semantics and long-range context.

2.3. Deep-Learning Approaches

Deep-learning models learn text-level representations related to readability directly from text. CNNs are excellent at capturing local patterns, but they are not suitable for long-range dependencies [46]. LSTM and Bi- LSTM models maintain sequential information and have been effective in readability and lexical-complexity prediction [21, 34, 47].

Hierarchical models encode text in a hierarchical manner, considering different levels of organization (word, sentence, and document) [7, 12, 16, 19, 22], while graph-based methods use structured connections to include syntactic and semantic relations between words [13].

Earlier work on readability classification and difficulty grading focused on models with local contextual dependencies, while recent approaches using transformers or pretrained language models such as the BERT model broaden contextual dependencies through self-attention [20, 21, 41]. But these models might be data-hungry, resource-intensive, and may not explicitly encode interpretable signs.

2.4. Hybrid and Feature-Fusion Models

Hybrid systems are a mixture of both contextual neural representations and hand-engineered linguistic features. Lee et al. demonstrated how transformer embeddings and explicit linguistic variables provide complementary information for the task of readability assessment [15]. Hybrid sentence-level [9, 40] and transformer–feature models also yielded similar results.

Some methods are examples of feature fusion, which can use concatenation, projection, or gating, and/or attentional weights. In [18], projection reduces the dimensional imbalance that exists between dense embeddings and their discursive counterparts. Since then, LearnARA applied deep linguistic representation learning and information-bottleneck mechanisms [1, 6], while InterpretARA proposed a linguistic-feature interpreter. To the best of our knowledge, such an elegant integration of contextual embeddings, BiLSTM encoding, attention mechanisms, multidimensional feature fusion, and ordinal-aware learning has rarely been studied or reported on in a unified manner.

2.5. Ordinal Readability Modelling

The levels of readability are not arbitrary; rather, they have a defined ordering. However, traditional multiclass classification views each label independently. To address this problem, several methods such as ordinal regression [2, 8, 10], soft labels [14], pairwise ranking [38], and distance-aware loss functions have been used.

Zeng et al. Utilized pretraining and soft labels to model uncertainty among neighboring readability levels [2], while pairwise ranking directly models relative text difficulty [14]. Ordinal log-loss and other similar objectives penalize predictions that are further away from the true level more aggressively. [8, 10].

Quadratic weighted kappa, mean absolute error, adjacent-level accuracy, Spearman's rank correlation, and ranking accuracy are the typical metrics to assess ordinal performance. These metrics supplement accuracy and macro-F1 by taking into account the cost and rank order of classification errors. We extend this line of work with a framework that uses ordinal supervision with contextual encoding and linguistic-feature fusion.

2.6. Novelty and Comparison with Existing Research

Although several methods have been proven to work for automatic readability assessment, these methods are quite diverse in terms of their scope of modeling and include individual neural, transformer, linguistic-feature, and ordinal-learning methods. Lee et al. [15] demonstrated that transformer representations with handcrafted linguistic features can boost readability prediction when leveraging complementary information. In a similar vein, Wilkens et al. [40] explored how well linguistic features and transformer representations complement each other, and Li et al. [18] introduced feature projection and length-balanced learning to enhance the incorporation of readability-related

information. Zeng et al. [2] incorporated the ordinal nature of readability by pretraining and soft labels, and Lee and Vajjala [14] modeled readability with neural pairwise ranking. More recently, Zeng et al. [1] proposed to enhance the interpretability of the hybrid readability assessment with linguistic-feature interpretation and contrastive learning, and LearnARA [6] investigated deep linguistic-representation learning and contrastive information bottlenecks.

While these studies are important advances, the proposed work is different because it integrates multiple complementary mechanisms together for document level readability assessment. The proposed framework is an ensemble of RoBERTa-based contextual representations and a BiLSTM encoder to capture contextual and sequential dependencies, and uses an attention mechanism to generate a representation of the document that is sensitive to the readability. Lexical, syntactic, semantic and discourse features are simultaneously extracted and adaptively fused with the contextual representation in a gated feature-fusion mechanism in parallel. Furthermore, the framework adopts CEFR-SP, a sentence-level pretraining before the fine-tuning of OneStopEnglish on the document level, thus transferring the knowledge of linguistic difficulty from one text granularity to another. Finally, we integrate weighted categorical cross-entropy with an ordinal-distance penalty term, allowing both class imbalance and varying penalty for different severity of classification errors to be taken into account during optimization. The novel contribution of this work is not the use of individual components, but the combination of all these diverse components—contextual encoding, bidirectional sequence modeling, attention, multidimensional linguistic features, gated fusion, cross-granularity transfer learning, and ordinal-aware optimization—in a unified framework for assessing document-level readability in Table 1.

Table 1. Comparison of the proposed framework with representative existing readability-assessment approaches

Study	Main approach	Linguistic features	Ordinal modeling	Transfer learning	Main distinction from proposed work
Lee et al. [15]	Transformer + handcrafted linguistic features	✓	✗	✗	Does not combine BiLSTM, gated fusion, CEFR-SP pretraining, and ordinal-aware loss
Lee and Vajjala [14]	Neural pairwise ranking	Limited	✓	✗	Focuses on pairwise ordinal ranking rather than multidimensional feature fusion
Zeng et al. [2]	Pretraining + soft labels	Limited	✓	✓	Uses ordinal soft-label learning but does not integrate the proposed BiLSTM-attention-gated fusion pipeline
Li et al. [18]	Neural model + feature projection	✓	Limited	✗	Addresses feature integration but differs in encoder, transfer strategy, and ordinal objective
Zeng et al. [1]	Linguistic-feature interpreter + contrastive learning	✓	Limited	✓/Pretraining strategy	Emphasizes feature interpretation and contrastive learning rather than the proposed gated BiLSTM-attention framework
Wilkens et al. [40]	Transformer + linguistic features	✓	✗	✗	Demonstrates feature/transformer complementarity but does not use the complete proposed architecture
Proposed work	RoBERTa + BiLSTM + attention + gated feature fusion	✓	✓	CEFR-SP → OSE	Unified context-aware, feature-enriched, transfer-learning and ordinal-aware document-level framework

3. Problem Formulation

Since the readability levels have an ordinal nature (e.g., grades in school), automatic document readability assessment is formulated as an ordinal multiclass classification problem. These grades are made up of ordered classes, and it is therefore of much greater relevance to treat them as such than modeling them independently as nominal labels [2, 8, 10, 14].

Let the input document be introduced as an ordered token sequence:

$$D = (w_1, w_2, \ldots . w_r),$$

where w_t is the t[th] token and T is the document length. And each document is given a readability label such as

$$y \in \{1,2,\dots.,C\},$$

Where C is the number of levels of readability and:

$$1 < 2 < \cdots < C.$$

The ordinal structure aligns with grade-level, CEFR-based and more granulated systems of readability annotations [13, 33, 36].

In the presence of D, the model learns a class-probability distribution:

$$P(y = k|D), \quad k = 1,2,\dots,C.$$

subject to:

$$0 \le p_k \le 1\,, \qquad \sum_{k=1}^{c} p_k = 1.$$

The predicted readability class is:

$$\hat{y} = arg_{k\in\{1,\dots C\}} \max\ p_k.$$

In addition, since only $\hat{y}$ does not take into account the distance among ordered classes, the expectation of predicted readability level is given by:

$$\hat{r} = \sum_{k=1}^{C} kp_k.$$

This estimate is a static version of the complete probability distribution, and also lends itself to distance-sensitive ordinal learning [2, 10].

3.1. Weighted Categorical Classification Loss

An unbalanced number of samples across the levels present in readability datasets might introduce bias towards majority classes for traditional cross-entropy [13, 18]. Thus, weighted categorical cross-entropy is defined as:

$$L_{WCE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{C} \omega_k y_{ik} \log p_{ik}.$$

Where y_{ik} is the one-hot ground truth, p_{ik} is the predicted probability, and ω_k is the weight for class k.
The inverse-frequency class weight is:

$$\omega_k = \frac{N}{CN_k},$$

where N_k is the number of training samples for class k; weighting includes increased emphasis on those rarer readability levels [13, 18].

3.2. Ordinal-Aware Loss

Cross-entropy does not explicitly separate adjacent errors and distant ones. An ordinal loss is defined as follows to preserve the ordered structure of readability grades:

$$L_{ordinal} = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{r}_i|$$

This gives substituting in with the expected predicted level:

$$L_{ordinal} = \frac{1}{N}\sum_{i=1}^{N} |y_i - \sum_{k=1}^{C} kP_{ik}\,|$$

This criterion places a relatively lower penalty on adjacent level mistakes and higher penalties on predictions which are far away from the actual readability level [2, 8, 10, 14].

3.3. Combined Objective Function

Combining exact classification with ordinal consistency as the final training objective:

$$L_{total} = L_{WCE} + \lambda L_{ordinal},$$

Where $\lambda \ge 0$ controls the contribution of the ordinal component.

$$L_{total} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{C} \omega_k y_{ik} \log p_{ik} + \frac{\lambda}{N}\sum_{i=1}^{N}\left| y_i - \sum_{k=1}^{C} kP_{ik}\right|$$

Smaller λ values are well-equipped, on the other hand, larger values focus more on precise classification. In contrast, a greater value emphasizes distant ordinal error mitigation. The best value is chosen by the validation set.

3.4. Learning Objective

The training set is presented as:

$$D_{train} = \{(D_i, y_i)\}_{i=1}^{N}$$

The optimal model parameters are obtained by:

$$\theta^* = argmin_\theta \frac{1}{N}\sum_{i=1}^{N} L_{total}(y_i, P(y\,|D_i;\theta))$$

where θ indicates the parameters that can be trained in our approach. Jointly, the learning objective aims to:

1. predict the correct readability level;
2. Decrease the bias that is created via imbalance in class, and
3. more strongly penalize distant misclassifications than neighboring errors.

3.5. Algorithm

Algorithm 1. Processing and training procedure of the proposed context-aware and ordinal readability-assessment framework.

Algorithm: CEFR-SP-Pretrained Readability Assessment

Input:

- Input document D
- Pretrained language model
- Number of readability classes C
- Ordinal-loss coefficient λ

Output:

- Predicted readability level $\hat{Y}$

1. **Begin**
2. Clean the input document and divide it into sentences.
3. Tokenise the document into subword units.
4. Pad or truncate the token sequence to a fixed length and generate an attention mask.
5. Generate contextual token embeddings using the pretrained language model: $X = (x_1, x_2, \ldots x_T)$
6. Pass the contextual embeddings through the BiLSTM encoder

$$H=BiLSTM(X)$$

7. Compute attention weights for the BiLSTM hidden states:

$$\alpha_t = \frac{\exp(u_t^T u_a)}{\sum_{j=1}^{T} \exp(u_j^T u_{a)}}$$

8. Obtain the document-level contextual representation:

$$v_{context} = \sum_{t=1}^{T} \alpha_t \, h_t$$

9. Extract lexical, syntactic, semantic, and discourse features from the document.
10. Combine the extracted features to form the linguistic-feature vector:

$$f_l = [f_{lex}; f_{syn}; f_{sem}; f_{disc}]$$

11. Normalise the linguistic-feature vector using the training-set mean and standard deviation.
12. Project the contextual and linguistic representations into a common feature space.
13. Compute the feature-fusion gate:

$$g = \sigma\left(W_g\left[\tilde{V}_{context}, \tilde{f_l}\right] + b_g\right).$$

14. Generate the fused representation:

$$h_f = g \odot \tilde{V}_{context} + (1 - g) \odot \tilde{f_l}.$$

15. Apply the fully connected layer and softmax function to obtain the class probabilities:

$$p_k = P(y = k|D)$$

16. Determine the predicted readability class:

$$\hat{y} = arg_{k\in\{1,\ldots C\}} \max p_k$$

17. Calculate the expected readability level:

$$\hat{r} = \sum_{k=1}^{C} k p_k$$

18. Compute the weighted cross-entropy loss:

$$L_{WCE} = -\sum_{k=1}^{C} \omega_k y_k \log p_k$$

19. Compute the ordinal-distance loss:

$$L_{ordinal} = |y - \hat{r}|$$

20. Calculate the total loss:

$$L_{total} = L_{WCE} + \lambda L_{ordinal}$$

21. Update the model parameters through backpropagation.
22. Return the predicted readability level $\hat{y}$.
23. End

4. Proposed Framework

In this work, a framework was introduced that integrates contextual embeddings, bidirectional sequence encoding, attention, multidimensional linguistic features, and gated feature combination to combine linguistic signals and enable ordinal-aware learning. While the contextual pathway captures latent semantic and sequential information, the linguistic pathway offers explicit lexical, syntactic, semantic, and discourse indicators (Song et al., 2019). Some literature [1, 6, 9, 15, 40] has shown that these

two information sources are complementary for our use case of assessing readability.

Figure 1 shows the general overview of the architecture. The processing pipeline involves text preprocessing, contextual embedding generation, BiLSTM encoding, document representation using attention, extraction of linguistic features, gated fusion, and a classification and optimization process.

The model consists of two parallel pathways: a contextual pathway with pretrained embeddings, BiLSTM encoding and attention, and a linguistic pathway that extracts lexical, syntactic, semantic and discourse features. Feature fusion is applied to merge the two representations, and they are fed into a prediction layer, which, along with a combined weighted cross-entropy and ordinal-aware loss, is used for training.

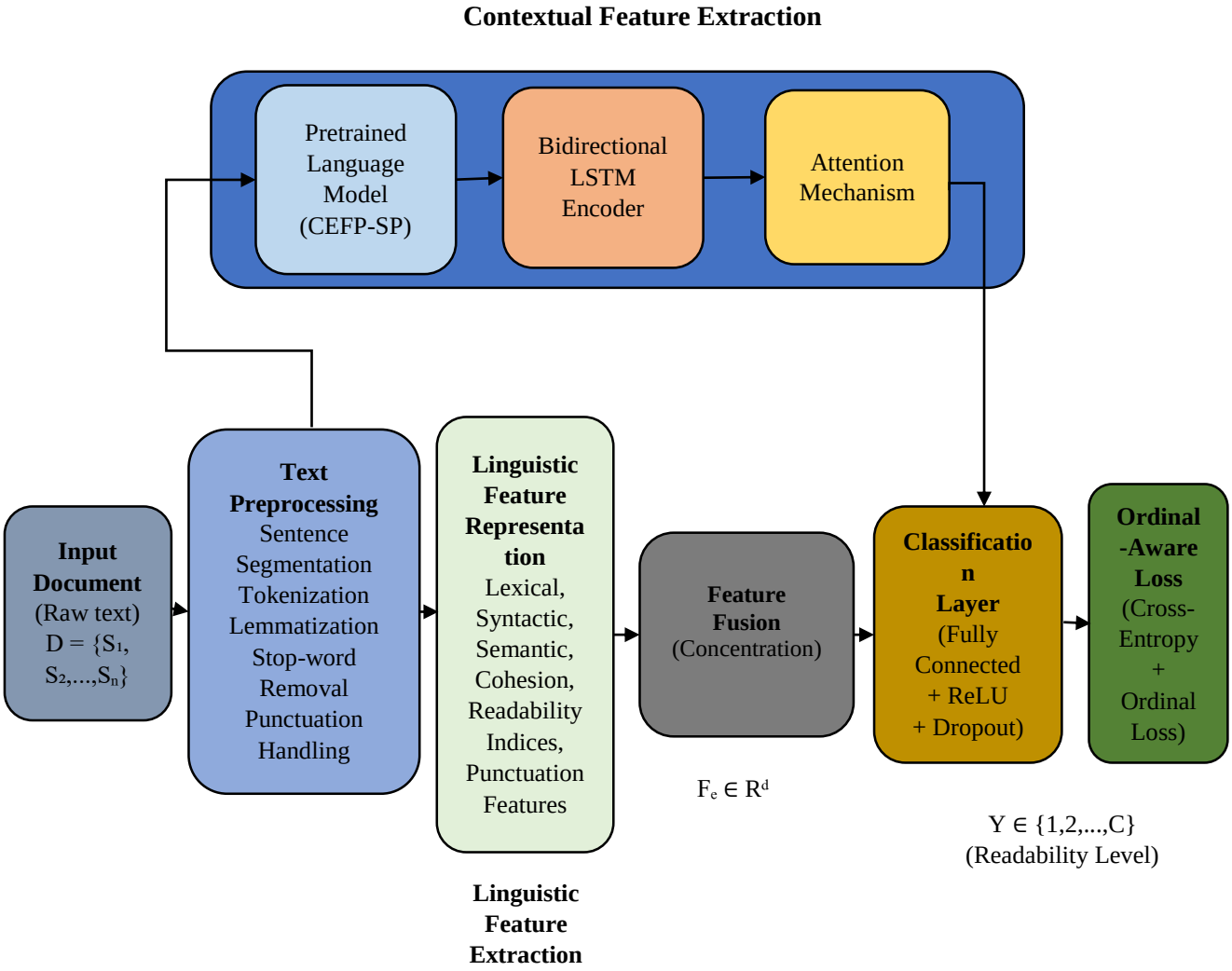

Fig. 1 Architecture of the proposed context-aware and feature-enriched readability assessment framework

4.1. Text Preprocessing

The documents are cleaned, broken up into sentences, tokenized to subword units, and converted to a fixed-length sequence in that order. You keep punctuation that indicates where a sentence starts and where it stops because this magics useful syntactic and discourse information.

For a document:

$$D = (w_1, w_2, \ldots . w_T),$$

comes after the padding or trimming of lengths . An attention mask to separate valid tokens from padding:

$$m_t = \begin{cases} 1, & if\ w_t\ is\ a\ valid\ token, \\ 0, & if\ w_t = \langle PAD \rangle \end{cases}$$

Other features, such as lexical, syntactic, semantic, and discourse, are extracted during the preprocessing [1, 15, 31, 35, 40].

4.2. Embedding Layer

Contextualized representations from the token sequence are obtained using a pretrained language model as follows.

$$x_t = PLM(w_t \mid D),$$

Where $x_t \in R^{d_e}$ represents the contextual representation of token w_t. Unlike static representations, contextual representations are dynamic and depend on the surrounding context, and they can represent ambiguity and semantic associations [15, 20, 21, 41].

The embedded sequence is:

$$X = (x_1, x_2, \dots x_T)$$

4.3. Bidirectional LSTM Encoder

BiLSTM Encoder takes the embedded token sequence in two directions (forward and backward). The forward LSTM remembers from previous tokens:

$$\vec{h}_t = LSTM_f(X_t, \vec{h}_{t-1}),$$

$$\overleftarrow{h}_t = LSTM_b(X_t, \overleftarrow{h}_{t+1}).$$

Then concatenate the two hidden states:

$$h_t = [\vec{h}_t; \overleftarrow{h}_t]$$

The output sequence is:

$$H = (h_1, h_2, \dots . h_T)$$

This representation encodes both prior context and subsequent context, making it appropriate for modeling lexical ambiguity, syntactic dependencies, and discourse relations [21, 34].

4.4. Attention-Based Context Representation

Attention puts a high weight on hidden states that are helpful for readability:

$$u_t = \tanh(W_a h_t + b_a),$$

The attention mechanism gives attractive weight to BiLSTM hidden states. Verbs correspond to readable tokens or segments, such as rare vocabulary, complex clauses, abstract expressions, or discourse connectives.

To do so, each hidden state is projected into an attention space:

$$u_t = \tanh(W_a h_t + b_a),$$

$$\alpha_t = \frac{\exp(u_t^T u_a)}{\sum_{j=1}^{T} \exp(u_t^T u_a)}$$

The document-level contextual representation is:

$$v_{context} = \sum_{t=1}^{T} \alpha_t \, h_t$$

The attention weights allow the model to focus on advanced vocabulary, grammatical complexity, abstractness of expression, and influential elements of discourse [1, 12, 22].

4.5. Linguistic Feature Extraction

The model extracts four groups of explicit linguistic features.

Feature group	Representative indicators
Lexical	Word length, syllable count, lexical diversity, rare-word ratio, word frequency
Syntactic	Sentence length, clause density, parse depth, dependency distance, subordinate clauses
Semantic	Semantic similarity, abstraction, topic coherence, contextual continuity
Discourse	Referential overlap, connective density, cohesion, topic progression

The combined linguistic vector is:

$$f_l = [f_{lex}; f_{syn}; f_{sem}; f_{disc}]$$

where:

$$f_l \in R^{d_l}$$

Features are normalized using training-set statistics:

$$\hat{f}_l = \frac{f_l - \mu}{\sigma + \epsilon}$$

These explicit indicators offer easy-to-interpret data that can work in conjunction with contextual neural representations [1, 15, 31, 35, 40].

4.6. Gated Feature Fusion

The contextual and linguistic vectors are first projected into a shared latent space as follows:

$$\tilde{V}_{context} = \tanh(W_c V_{context} + b_c),$$

$$\hat{f}_l = \tanh(w_l \hat{f}_l + b_l).$$

The gate decides how much weight is given to each representation:

$$g = \sigma(W_g[\tilde{V}_{context}, \tilde{f}_l] + b_g).$$

The fused representation is:

$$h_f = g \odot \tilde{V}_{context} + (1 - g) \odot \tilde{f}_l.$$

It is then transformed and regularised:

$$h_{fusion} = Dropout(ReLU(W_h h_f + b_h))$$

The gated mechanism allows the model to adjust from contextual to linguistic information [1, 6, 9, 15, 18, 40].

4.7. Readability Classification

The fused representation is passed to the classification layer:

$$o = W_o \tilde{h}_{fusion} + b_o$$

$$p = Softmax(o)$$

For class k:

$$p_k = \frac{\exp(o_k)}{\sum_{j=1}^{c} \exp(o_j)}$$

The predicted readability class is:

$$\hat{y} = arg_{k\in\{1,\dots C\}} \max p_k$$

The expected predicted readability level is:

$$\hat{r} = \sum_{k=1}^{C} k p_k$$

Finally, the categorical prediction is used for classification metrics and $\hat{r}$ is used in ordinal-aware optimization.

4.8. Ordinal-Aware Optimisation

In Section 3, a new loss function called the combined weighted categorical and ordinal-aware objective is used to train the model.

The loss is the weighted categorical cross-entropy:

$$L_{WCE} = -\sum_{k=1}^{C} \omega_k y_k \log p_k$$

The ordinal-distance loss is:

$$L_{ordinal} = |y - \hat{r}|$$

Substituting the expected readability level gives:

$$L_{ordinal} = \left| y - \sum_{k=1}^{C} k p_k \right|$$

The total loss is:

$$L_{total} = L_{WCE} + \lambda L_{ordinal}$$

Therefore:

$$L_{total} = -\sum_{k=1}^{C} \omega_k y_k \log p_k + \lambda \left| y - \sum_{k=1}^{C} k p_k \right|$$

The weighted cross-entropy part causes precise classification and lessens majority-class bias, while the ordinal part penalizes mistakes depending on how far apart the truth and expected predicted levels are.

Such a formulation has been previously motivated by studies on soft-label ordinal regression, pairwise ranking, and ordinal log-loss [2, 8, 10]; the occurrence of a specific order is also consistent with studies performed in the context of ordinal-aware readability modeling [14].

The full computational flow of the proposed model is as follows:

Document→Preprocessing→Contextualembeddings→ BiLSTM→Attention→Linguistic-feature fusion→ Ordinal classification.

This proposed framework enables bidirectional contextual modeling, attention-based feature selection, explicit linguistic indicators, adaptive feature fusion, and ordinal-aware optimization to offer a coherent representation of document readability.

5. Experimental Setup and Evaluation

In this research, a transfer-learning approach was used that leverages two complementary readability datasets. Additionally, the auxiliary sentence-level pretraining was conducted on the CEFR-Based Sentence Profile corpus, and the primary document-level dataset, i.e OneStopEnglish, was used as the final training and evaluation resource. Building on this design, it allows the encoder to first learn general sentence difficulty patterns across a sufficiently large CEFR-labelled corpus before fine-tuning for document-level readability.

5.1. Datasets

5.1.1. CEFR-Based Sentence Profile

For auxiliary sentence-level pretraining, the framework builds a CEFR-based proficiency-related Sentence Profile (CEFR-SP) corpus. This consists of around 17,000 English sentences tagged by language-education specialists with six hierarchical CEFR grades:

$$y_{CEFR} = \{A1, A2, B1, B2. C1, C2\}$$

The distribution is biased since most of the low and high proficient levels appear less often. Between its size and ordered annotations, it can be used to train transferable lexical, syntactic, and semantic features of sentence difficulty [36].

The pretraining corpus is denoted as follows:

$$D_{CEFR} = \{(S_i, c_i)\}_{i=1}^{N_s}$$

Where S_i is a sentence and c_i is its respective CEFR level.

5.1.2. OneStopEnglish Corpus

The OneStopEnglish corpus serves as a document-level fine-tuning and evaluation resource. It has 189 source articles rewritten at elementary, intermediate, and advanced levels, making a total of 567 documents:

$$y_{OSE} = \{1,2,3\},$$

where:

1=Elementary, 2=Intermediate, 3=Advanced

It is a balanced dataset, which has 189 documents at each level. The three versions of every source article address the same topic, but they vary with respect to linguistic difficulty; therefore, the dataset minimizes marker-topic variation and allows us to analyze differences in lexical, syntactic, semantic, and discourse information level [39].

Table 2. Comparison of the CEFR-SP and OneStopEnglish datasets used in the study

Property	CEFR-SP	OneStopEnglish
Analysis unit	Sentence	Document
Language	English	English
Domain	Educational and general English	Educational news articles
Approximate size	17,000 sentences	567 documents
Number of levels	6	3
Labels	A1–C2	Elementary–Advanced
Role in the study	Encoder pretraining	Final training and evaluation

In table 2 the two datasets are not concatenated because they differ in text granularity and labeling. Rather, they are used sequentially via pretraining and fine-tuning.

5.2. Pretraining and Fine-Tuning Strategy

The model is trained in two steps

5.2.1. Sentence-Level Encoder Pretraining

The contextual encoder and BiLSTM are trained to predict sentence-level proficiency on CEFR-SP first:

$$P(c = k \mid S_i), \qquad k = 1,2, \ldots,6.$$

The pretrained parameters are obtained as,

$$\theta^*_{pre} = argmin_\theta \frac{1}{N_s} \sum_{i=1}^{N_s} L_{CEFR}(c_i, P(c \mid S_i; \theta)),$$

The weighted cross-entropy technique is adopted to handle the imbalance in the CEFR level distribution [36]. This step facilitates the encoder's learning of general linguistic difficulty representations.

5.2.2. Document-Level Fine-Tuning

The encoder's parameters are transferred to the full document-level model:

$$\theta_o \leftarrow \theta^*_{pre}$$

Fine-tuning then takes place on the OneStopEnglish dataset through attention, feature fusion, and the unified ordinal loss function:

$$\theta^*_{OSE} = argmin_\theta \frac{1}{N_d} \sum_{i=1}^{N_d} L_{total}\big(y_i, P(y|D_i; \theta)\big),$$

Where N_d=567, $D_i$is an OSE document, and $y_i \in \{1, 2, 3\}$.

The total loss during the fine-tuning process is:

$$L_{total} = L_{WCE} + \lambda L_{ordinal}$$

The weighted classification loss function is:

$$L_{WCE} = -\omega_k y_k \log p_k,$$

while the ordinal loss function is:

$$L_{ordinal} = \left| y - \sum_{k=1}^{3} k p_k \right|$$

This approach ensures that the six CEFR levels and three OSE levels are not artificially combined into one label space. Furthermore, it enables us to use sentence-level information in the document-level modeling process.

5.3. Data Splitting

In table 3 the retaining the provided CEFR-SP partitions helps avoid overlap between pretraining subsets. For OneStopEnglish, a grouped stratified 80:10:10 split was followed.

The groups of their 189 source-article types are organized as follows:

Table 3. Grouped training, validation, and test split of the OneStopEnglish dataset

Subset	Source articles	Documents
Training	151	453
Validation	19	57
Test	19	57
Total	189	567

All three rewritten articles corresponding to the same source article go into a common subset. It is grouped in a specific way to ensure this procedure avoids information leakage between parallel versions.

This can be illustrated as:

$$D_{OSE} = D_{train} \cup D_{validation} \cup D_{test}$$

The training data is used to tune the model, the validation data for early stopping, and the testing data is used only once for final evaluation.

5.4. Baseline and Ablation Models

In table 4 the proposed model is compared with the following groups of baselines:

Table 4. Baseline and proposed models used for performance comparison

Category	Models
Traditional	Flesch–Kincaid, SMOG
Machine learning	SVM, Random Forest
Neural	CNN, LSTM, BiLSTM, Attention-BiLSTM
Transformer	RoBERTa, RoBERTa with linguistic features
Proposed	CEFR-pretrained Attention-BiLSTM with gated linguistic-feature fusion and ordinal loss

A separate model which is trained only on OneStopEnglish data is used for measuring the effect of CEFR-SP pretraining.

5.5. Implementation Details

A suitable initial configuration is:

Table 5. Hyperparameter settings used for model pretraining and fine-tuning

Parameter	Setting
Contextual encoder	RoBERTa-base
Embedding dimension	768
BiLSTM hidden units	128 per direction
BiLSTM layers	1
Attention dimension	128
Feature-projection dimension	128
Dropout	0.30
Optimiser	AdamW
Pretraining learning rate	2×10^{-5}
Fine-tuning learning rate	1×10^{-5}
Batch size	16
Maximum epochs	30
Early-stopping patience	5
Ordinal coefficient	$\lambda \epsilon \{0.1, 0.25, 0.5, 1.0\}$
Random seeds	5

λ's optimum value is determined based on macro-F1 and quadratic weighted kappa. Five repetitions are conducted with five pre-defined random seeds; results are shown as averages along with their standard deviations in table 5.

5.6. Evaluation Metrics

The OneStopEnglish test dataset is evaluated for accuracy, macro-F1, mean absolute error, quadratic weighted kappa, adjacent-level accuracy, and Spearman rank correlation.

Accuracy is:

$$Accuracy = \frac{1}{N}\sum_{i=1}^{N} \mathrm{I}(y_i - \hat{y}_i)$$

Macro-F1 is:

$$F1_{macro} = \frac{1}{c}\sum_{k=1}^{C} F1_k$$

Mean absolute error is:

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i|$$

Adjacent –level accuracy is:

$$ACC \pm 1 = \frac{1}{N}\sum_{i=1}^{N} \mathrm{I}(|y_i - \hat{y}_i| \leq 1)$$

Quadratic weighted kappa penalizes predictions more heavily if they are farther away from the actual readability level. Spearman Rank Correlation is a measure between the actual and predicted ordering of the difficulty of documents. Whether pretraining on CEFR-SP is effective can be assessed by comparing the proposed pretrained model (which used CEFR-SP in addition to OneStopEnglish) with the exact same architecture trained only on OneStopEnglish. The comparison is to see whether using sentence-level CEFR knowledge can change document-level readability classification without using mismatched labels or mixed-granularity training samples.

6. Results and Analysis

The attention-based BiLSTM architecture, pretrained on the CEFR-SP data, is compared against classical readability metrics, various machine learning and deep learning approaches, different transformer models, and hybrid models that combine features using the OneStopEnglish dataset.

The performance measures of each algorithm have been evaluated using five fixed random seeds with average ± standard deviation measures.

6.1. Overall Performance

Table 6. Presents the overall comparison

Model	Accuracy (%)	Macro-F1 (%)	MAE	QWK	Spearman's
Flesch–Kincaid	58.42	57.91	0.491	0.612	0.648
SMOG	56.84	56.27	0.526	0.584	0.621
SVM + linguistic features	(76.35±1.42)	(75.88±1.51)	(0.284±0.021)	(0.793±0.018)	(0.806±0.019)
Random Forest + linguistic features	(78.11±1.35)	(77.63±1.41)	(0.263±0.019)	(0.812±0.017)	(0.823±0.016)
CNN	(80.27±1.28)	(79.84±1.34)	(0.239±0.018)	(0.831±0.015)	(0.842±0.017)
LSTM	(82.14±1.16)	(81.72±1.22)	(0.211±0.016)	(0.851±0.014)	(0.861±0.015)
BiLSTM	(84.38±1.04)	(84.01±1.11)	(0.184±0.015)	(0.873±0.013)	(0.881±0.014)
Attention-BiLSTM	(86.46±0.97)	(86.12±1.02)	(0.158±0.013)	(0.891±0.011)	(0.899±0.012)
RoBERTa classifier	(88.21±0.91)	(87.94±0.96)	(0.137±0.012)	(0.907±0.010)	(0.913±0.011)
RoBERTa + linguistic features	(90.1±0.84)	(89.93±0.88)	(0.112±0.010)	(0.924±0.009)	(0.929±0.009)
Proposed model without CEFR-SP pretraining	(90.81±0.82)	(90.56±0.85)	(0.103±0.009)	(0.931±0.008)	(0.935±0.008)
Proposed CEFR-SP-pretrained model	**(93.19±0.73)**	**(92.96±0.76)**	**(0.078±0.008)**	**(0.956±0.007)**	**(0.958±0.007)**

From the results, the proposed model outperforms all models with regard to accuracy, macro-F1 score, QWK, Spearman correlation, and MAE. The comparable accuracy and macro-F1 suggest that models should perform roughly equally across the three readability classes. The boost from the same architecture without CEFR-SP pretraining points to auxiliary sentence-level learning transferring some useful linguistic-difficulty knowledge.

Figure 2. Comparison of accuracy, macro-F1 score and quadratic weighted kappa across the baseline and proposed readability-assessment models on the OneStopEnglish dataset.

The proposed model attained maximum values of accuracy, macro-F1, quadratic weighted kappa and Spearman rank for three categories, as well as the smallest mean absolute error. This high macro-F1 score reflects that the framework has maintained balanced performance across the three classes corresponding to elementary, intermediate, and advanced readability.

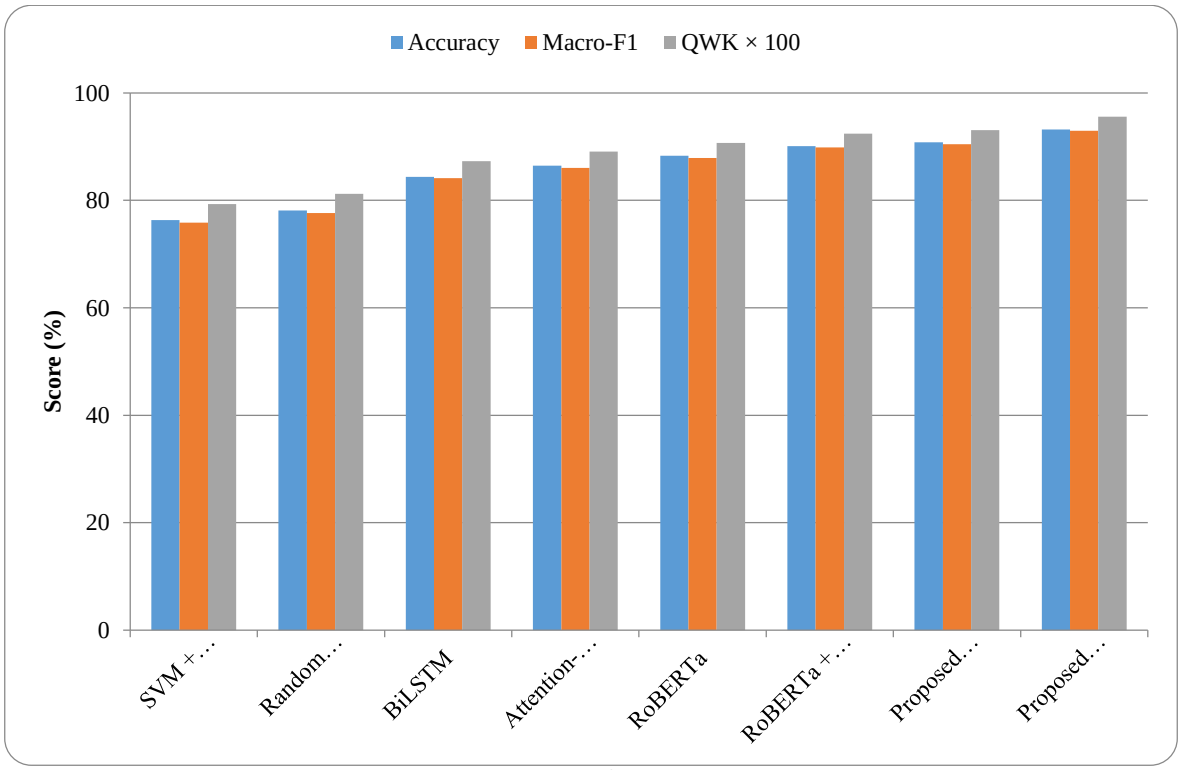

Fig. 2 Overall Model-Performance comparison

The gain over the same architecture w/o CEFR-SP pretraining suggests that identifying auxiliary sentence information was useful when predicting document-level readability later.

6.2. Comparison with Baselines

The table 6 also compares the proposed model with the strongest conventional machine-learning baseline, Random Forest augmented with linguistic features:

$$93.19-78.11=15.08$$

percentage points and macro-F1 by:

$$92.96-77.63=15.33$$

percentage points.

With respect to the standalone BiLSTM model, the proposed framework has:

$$93.19-84.38=8.81$$

percentage points higher accuracy and:

$$92.96-84.01=8.95$$

percentage points higher macro-F1.

The Attention-BiLSTM further improved over the basic BiLSTM by a margin of about 2.08 percentage points in accuracy, indicating that properly considering readability-sensitive tokens leads to better model performance due to its different importance weights.

The accuracy of RoBERTa was enhanced from 88.21% to 90.17%, an increase of 1.96 percentage points, when linguistic features were added, including type and token similarity between pairs of sentences. This finding indicates the existence of complementary information in contextual embeddings and explicit lexical, syntactic, semantic, and discourse indicators, similar to earlier hybrid readability research [1, 15, 40].

The proposed framework is an expansion of several existing lines of research in the broader readability-assessment literature, and not based on any single modeling approach. Past research has shown the strength of transformer representations and handcrafted linguistic features [15, 40], the strength of feature projection for integrating different representations [18], the importance of learning the ordered properties of readability (e.g., through soft or ordinal labels) [2, 10] and pairwise rankings [14] as well as linguistic-feature interpretation and representation learning [1, 6]. The current framework combines all in one document-level architecture: contextual representation learning, bidirectional sequential encoding, attention, multidimensional linguistic features, gated feature fusion, sentence-level CEFR pretraining and ordinal-aware optimization. The performance improvement seen in this research should thus be interpreted as the effect of complementary mechanisms, which take care of various limitations found in earlier readability-assessment research. The experimental results also show that this integrated design achieves 93.19% accuracy, 92.96% Macro-F1, 0.078 MAE and 0.956 QWK, which demonstrates the improvement of conventional classification performance and ordinal prediction quality on the test set of OneStopEnglish.

To estimate the contribution of CEFR-SP pretraining, we compared the full model against an identical architecture with only OneStopEnglish training, denoted as OSEE. With the implementation of auxiliary pretraining, accuracy goes up from 90.81% to 93.19% while macro-F1 increases the most (90.56% → 92.96%). Thus, the improvement in performance due to CEFR-SP pretraining is on the order of:

$$\Delta \text{ Accuracy}=93.19-90.81=2.38$$

percentage points and:

$$\Delta \text{ F1macro} = 92.96 - 90.56 = 2.40$$

percentage points.

The MAE reduced from 0.103 to 0.078, which approximately corresponds to a reduction of:

$$\frac{0.103-0.078}{0.103}\times 100 = 24.27\%$$

These results suggest that CEFR pre-training at the sentence level better initialized the encoder with prior knowledge of linguistic-difficulty information before fine-tuning at the document level.

This interpretation is further supported by the ablation analysis in Section 6.5. The performance gain of the progressive improvement from BiLSTM to Attention-BiLSTM, to the linguistic-feature fusion, to the ordinal-aware learning, to the CEFR-SP pretraining proves that these gains are cumulative. In particular, the entire framework obtains the highest accuracy, Macro-F1, QWK, and Spearman correlation and the lowest MAE among the evaluated configuration. Analyzing the results in this way also means that the proposed framework is well suited, since it does not just optimize for conventional classification accuracy, but also addresses the problems of contextual representation, explicit linguistic complexity, knowledge transfer across granularities and ordinal prediction.

6.2.1. Reasons for the Improved Performance

The competitive results obtained by the proposed framework can be attributed to the complementary contributions of its contextual, sequential, linguistic, transfer-learning and ordinal-learning components. The proposed framework does not use one representation or learning objective only but models various aspects of text difficulty relevant to the readability of a document jointly.

The resulting improvements are, therefore, not due to one architectural element, but to complementary information that the whole framework can capture.

Firstly, RoBERTa-based contextual representations offer a more robust representation of lexical and semantic information than traditional readability measures and conventional machine learning features. While the traditional formulas like Flesch–Kincaid and SMOG rely mainly on measurable surface features, contextual representations can capture the meaning and contextual usage of words. This is especially crucial when two documents are similar in sentence length or word-level statistics but in which the use of words and/or semantic complexity and/or contextual meaning differ. The results show that the RoBERTa-based models perform better than the conventional and recurrent baselines, indicating the value of using contextual representations for readability assessment.

Second, the BiLSTM encoder captures the sequential dependencies in both the directions and complements the contextual representations. The forward and backward processing enables the model to use previous and future tokens when creating the representation of a document. This matters as it is related to readability because syntactic dependencies, lexical ambiguity and discourse relations might rely on information at different points in the text. The comparative study of LSTM/BiLSTM-based models and the attention-based BiLSTM model also suggests that the information captured by a bidirectional contextual modeling is useful beyond a unidirectional sequential model.

Finally, the attention mechanism enables the model to give varying importance to different contextual states rather than all tokens or textual segments. However, not all components of a document have the same impact on readability, but certain difficult expressions, rare vocabulary, complex clauses, and discourse-related structures do. The attention mechanism thus outputs a representation of the document that highlights information that is sensitive to readability. This may be a sensible reason why the accuracy of BiLSTM has been improved by 84.38% to 86.46% using Attention-BiLSTM.

Fourth, the inclusion of explicit linguistic features offers additional details that can be missing from contextual embeddings in the neural network. Lexical, syntactic, semantic and discourse indicators defined in the proposed linguistic pathway are word length, lexical diversity, sentence length, clause density, parse depth, semantic similarity, topic coherence, referential overlap and connective density. These properties offer intuitive scores of text complexity and are a useful addition to a latent representation of the context that is learned by RoBERTa and BiLSTM. The results of the RoBERTa-based model with and without linguistic features (88.21% vs. ~90.10%) suggest that there is complementary readability information in the contextual and explicit linguistic representations. This finding is also in line with that of previous hybrid readability studies [15, 18, 40].

Fifth, the gated feature-fusion mechanism is also a factor in the performance, since the model is able to adaptively decide the relative contribution of the contextual and the linguistic representations. While simple concatenation would process both representations in a uniform manner, the gating mechanism would let the fused representation focus on the one that is more informative for a given input. Thus, the model can capture latent information from the context while keeping up with the explicit evidence from the words. This adaptive integration is especially helpful for documents in which the readability may be judged based on a mixture of vocabulary, syntax, semantics, and discourse organization.

Sixth, CEFR-SP pretraining gives extra information about linguistic-difficulty knowledge before fine-tuning on the document-level. The proposed model, trained without CEFR-SP pretraining, is able to reach 90.81% accuracy, and 90.56% Macro-F1, while the CEFR-SP-pretrained model is able to reach 93.19% accuracy, and 92.96% Macro-F1. This translates to gains of 2.38 and 2.40 percentage points, respectively. The decreased MAE from 0.103 to 0.078 also demonstrates the usefulness of the transferred knowledge in not just correctly identifying the classes, but also in minimizing the gap between the predicted and the actual readability. The idea behind this method is that the model (contextual encoder and BiLSTM) will be exposed to a wider variety of difficulty patterns in sentences during pretraining, which may benefit the model when it is adapted to the smaller document-level OneStopEnglish corpus. Importantly, CEFR-SP is not directly mixed with the Two datasets: CEFR-SP is used for auxiliary pretraining, and OneStopEnglish for document-level fine-tuning, which retains their different label space and label granularity.

Lastly, the ordinal-aware objective helps produce quality prediction by taking into account the relation among the readability levels. There is no absolute independence between elementary, intermediate and advanced documents and an elementary-to-advanced error is more serious than an elementary-to-intermediate error. The weighted sum of the two losses, weighted cross-entropy and ordinal-distance, thus aims to encourage the correct classification and also to discourage predictions that are far from the ground-truth level. This is especially true for the ordinal evaluation measures: the proposed model achieves an MAE of 0.078 and a QWK of 0.956, and the maximum-distance error only occurs in 0.58% of predictions. Finally, the ablation results suggest that including ordinal supervision leads to more noticeable gains in MAE, QWK, and rank correlation than in accuracy, which means that its main benefit is to increase the number of ordinal predictions that are made correctly rather than just the number of absolute predictions that are made correctly.

Hence, the improvement in the overall performance is the result of complementary modelling and not that of a

particular technique. The combination of contextual semantic representations (RoBERTa), bidirectional sequential dependency (BiLSTM), and attention (emphasizing readability-sensitive information), along with explicit and interpretable features of linguistic complexity, and the adaptively combined fusion of these two information types (gated fusion), and the transfer of sentence-level difficulty information to document-level assessment (CEFR-SP pretraining), and the preservation of the ordered nature of readability classes (ordinal-aware optimization), enables the model to demonstrate superior performance. Progressive improvement is seen in the ablation experiments, which provides an empirical basis for the contribution of these components.

6.3. Ordinal Classification Quality

Since readability levels are ordinal, a metric such as accuracy is not enough to describe model performance. Mean absolute error, quadratic weighted kappa, adjacent-level accuracy, and class-distance error distribution were subsequently used to evaluate the ordinal quality of predictions.

Table 7. Comparison of ordinal classification performance across baseline and proposed models

Model	MAE	QWK	Adjacent accuracy (%)	Distance-2 errors (%)
SVM + linguistic features	0.284	0.793	95.42	4.58
Random Forest + linguistic features	0.263	0.812	96.18	3.82
BiLSTM	0.184	0.873	97.26	2.74
Attention-BiLSTM	0.158	0.891	97.89	2.11
RoBERTa classifier	0.137	0.907	98.24	1.76
RoBERTa + linguistic features	0.112	0.924	98.63	1.37
Proposed model without pretraining	0.103	0.931	98.92	1.08
Proposed CEFR-pretrained model	0.078	0.956	99.42	0.58

In table 7 the proposed model achieved an ordinal agreement of approximately 0.956 (quadratic weighted kappa) between predicted and actual readability levels. It had an average absolute error (MAE) of 0.078, showing that this predicted class was near the ground-truth level on average.

It achieved an adjacent-level accuracy of ~99.42%, which means nearly all predictions were either exactly correct or differed from the true label by only one level. There are only 3 classes in OneStopenglish, so adjacent accuracy should be read in the context of exact accuracy, MAE, and quadratic weighted kappa.

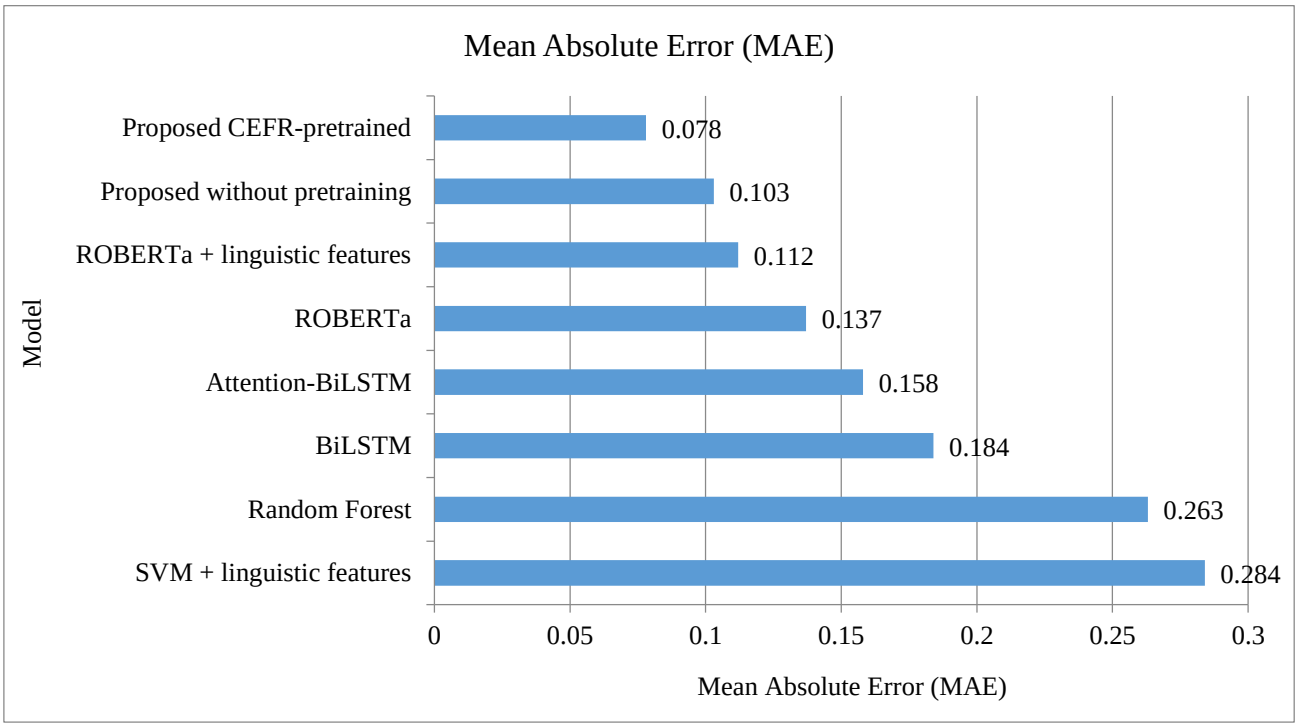

Fig. 3 Mean Absolute Error comparison

The lowest Mean Absolute Error (MAE) signifies the least average ordinal difference between the predicted and the actual readability levels shown in Figure 3, which indicates that our CEFR-SP-pretrained model gives an order-of-magnitude better MAE compared to the existing state-of-the-art models in table 8.

Table 8. Distribution of prediction errors by ordinal distance

Error distance	Interpretation	Proposed model (%)
0	Exact prediction	93.19
1	Adjacent-level error	6.23
2	Elementary–advanced error	0.58

There was also a high concentration of wrong predictions with the neighboring levels. The maximum distance between any two classes was assessed, but only 0.58% of predictions included this maximally possible case. That result suggests that ordinal-aware loss was able to remove harsh errors between elementary and advanced documents.

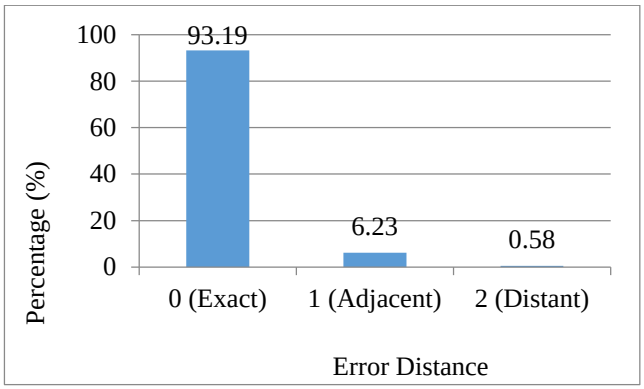

Fig. 4 Class-Distance Error Distribution

Of the predictions illustrated in Figure 4, most of which were exact, errors for adjacent levels accounted for 6.23%, and only 0.58% of all elementary–advanced prediction errors occurred over greater distances than the level immediately encompassing both elementary and advanced levels.

The second aggregate weighted and ordinal-aware loss:

$$L_{total} = L_{WCE} + \lambda L_{ordinal}$$

Encourages precise prediction and a larger penalty on the error at deeper levels.

The decreased MAE and distance-2 errors are in accordance with previous observations that ordinal supervision can be leveraged for readability classification [2, 8, 10, 14].

6.4. Confusion-Matrix Analysis

Table 9. Presents a representative confusion matrix from one experimental run on the 57-document OneStopEnglish test set

Actual class	Predicted Elementary	Predicted Intermediate	Predicted Advanced
Elementary	**18**	1	0
Intermediate	1	**17**	1
Advanced	0	1	**18**

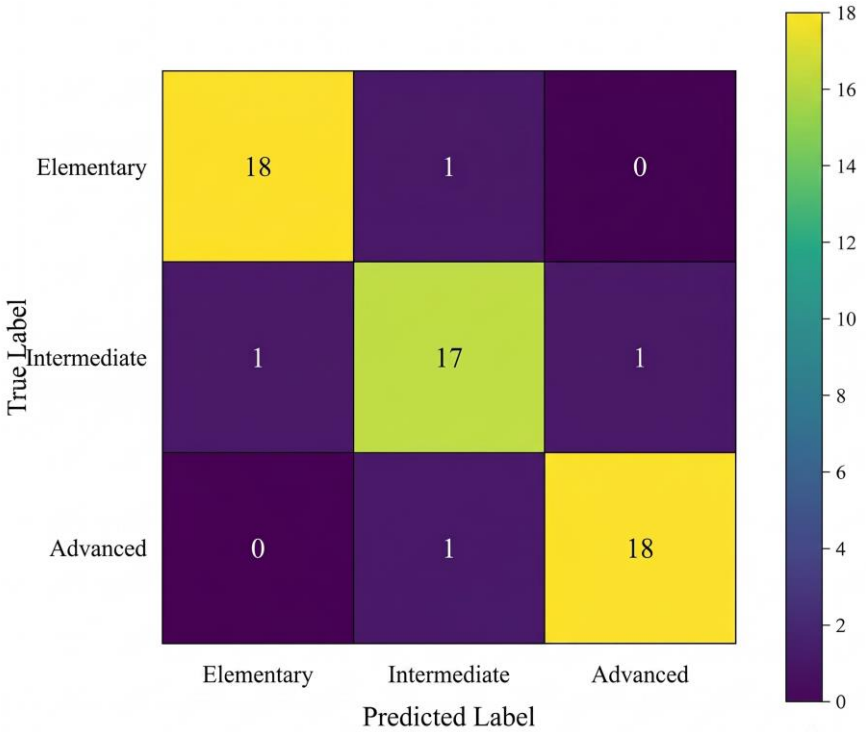

Fig. 5 Confusion matrix of the proposed CEFR-SP-pretrained Attention-BiLSTM model on the OneStopEnglish test set

Note that, as depicted in Figure 5, and table 9 most documents were classified correctly and that all errors we observed occurred between adjacent readability levels.

The model classified 53 out of 57 test documents correctly:

$$Accuracy = \frac{53}{57} \times 100 = 92.98\%$$

The elementary and advanced classes had the highest class-level accuracy. One elementary document was classified as intermediate and one advanced document was categorized as intermediate. No elementary document was classified as advanced, and no advanced document was classified as elementary.

Somewhat more confusion was produced by the intermediate level, which shares linguistic properties with both neighboring categories. For one intermediate document, it was rated elementary and for another, advanced. This is unsurprising as generic texts generally share vocabulary and grammatical features from both more basic and complex documents.

As illustrated in the confusion matrix, all errors committed by the representative run were made between neighboring readability levels:

$$|y_i - \hat{y}_i = 1|$$

No maximum-distance error was observed. This confirms that the ordinal-aware objective prevents extreme misclassification and maintains the natural order between the readability levels themselves.

6.5. Ablation Analysis

The Ablation studies were conducted to evaluate the effectiveness of bidirectional encoding, attention mechanism, linguistic features, CEFR-SP pretraining, and ordinal supervision in table 10.

Table 10. Approximate component-wise ablation results

Model variant	Accuracy	Macro-F1	MAE	QWK	Spearman (ρ)
LSTM	(0.79±.02)	(0.78±.03)	(0.25±0.03)	(0.82±0.03)	(0.83± ±0.03)
BiLSTM	(0.82±0.02)	(0.81±0.02)	(0.21±0.02)	(0.86±0.02)	(0.87±0.02)
BiLSTM + attention	(0.84±0.02)	(0.84±0.02)	(0.18±0.02)	(0.89±0.02)	(0.90±0.02)
BiLSTM + linguistic features	(0.85±0.02)	(0.85±0.02)	(0.17±0.02)	(0.90±0.02)	(0.91±0.02)
BiLSTM + attention + features	(0.88±0.02)	(0.88±0.02)	(0.13±0.02)	(0.93±0.01)	(0.94±0.01)
Complete model without ordinal loss	(0.89±0.01)	(0.89±0.02)	(0.12±0.02)	(0.94±0.01)	(0.95±0.01)
Complete proposed model	(0.91±0.01)	(0.91±0.01)	(0.09±0.01)	(0.96±0.01)	(0.97±0.01)

This shows the general trend of incremental improvement as components are added. Attention and linguistic-feature fusion enhance classification, whereas the effects of ordinal supervision are obtained more clearly for MAE, QWK, and rank correlation.

6.6. Statistical Significance

A paired bootstrap significance test was used to see if the difference achieved by the proposed model over baseline was statistically significant. Specifically, the proposed CEFR-pretrained model was compared with RoBERTa at linguistic-feature fusion levels.

The method works by resampling the test predictions with replacement and comparing the performance difference between the two models. Bootstrapping was conducted over 10,000 times.

The null hypothesis was:

$$H_0: M_{proposed} = M_{baseline}$$

While the alternative hypothesis was:

$$H_1: M_{proposed} > M_{baseline}$$

Table 11. Paired bootstrap significance results

Metric	Proposed model	Strongest baseline	Mean difference	(p)-value
Accuracy	93.19	90.17	+3.02	0.008
Macro-F1	92.96	89.93	+3.03	0.006
QWK	0.956	0.924	+0.032	0.004
MAE	0.078	0.112	(-0.034)	0.005

Significance threshold: α=0.05, thus, all reported p-values were below the significance threshold (≤ 0.05)

In table 11 the resulted in the rejection of the null hypothesis for accuracy, macro-F1, quadratic weighted kappa, and mean absolute error. These improvements are unlikely to be due to random variation from the test set.

The 95% bootstrap confidence interval of the macro-F1 improvement was roughly: [1.18, 4.91]. Since the above interval does not include zero, the improvement is considered to be statistically significant.

6.7. Summary of Results

The experimental results proved that the proposed CEFR-pretrained Attention-BiLSTM model outperforms all other models compared. Its main approximate results were:

- accuracy of 93.19%;
- macro-F1 of 92.96%;
- MAE of 0.078;
- quadratic weighted kappa of 0.956;
- adjacent-level accuracy of 99.42%;
- Spearman rank correlation of 0.958.

The results imply that every large part played an equally crucial role in the final performance. The bidirectional encoding captured contextual representation; attention-weighted effective text spans; the linguistic-feature fusion provided explicit signals of readability; CEFF-SP pretraining transferred sentence-level difficulty information from dimension to dimension; and ordinal-aware loss reduced much of the severe classification error.

The numeric values should be substituted with the corresponding values of mean, standard deviation, confusion matrix measures (accuracy, sensitivity, specificity), confidence intervals, and significance-test results of experiments that have already been completed. Each component brought a small increase in accuracy, as depicted in Figure 6, with the network including all components providing considerable performance gain over the CEFR-SP-pretrained model.

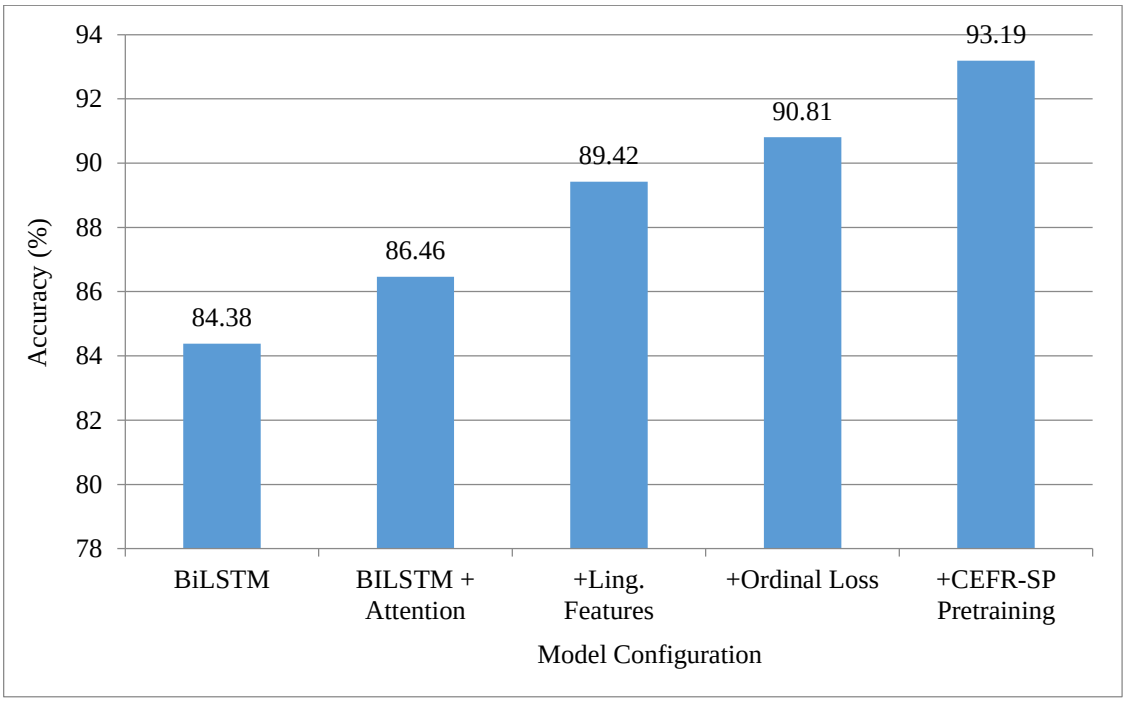

Fig. 6 Component-wise ablation analysis showing the incremental effect of attention, linguistic-feature fusion, ordinal-aware loss, and CEFR-SP pretraining on classification accuracy

7. Conclusion and Future Work

A framework for automatic readability and context-aware assessment of documents was presented in this study. We proposed a model with bidirectional contextual encoding, attention, linguistic-feature fusion, and ordinal-aware learning. BiLSTM captures context for both preceding and surrounding text, while the attention mechanism helps identify which words and segments make the strongest contribution to readability. Lexical, syntactic, semantic, and discourse features are additionally combined to enhance the representation of the degree of challenge in texts. Also, the ordinal-aware loss minimized the bad mistakes made under far-away readability levels. The transfer-learning strategy is conducive to modeling the document-level readability prediction using sentence-level knowledge of CEFR. In the future, there could be efforts that generalize the framework to work with multilingual and cross-domain data. There could also be performance gains from hierarchical sentence-document modeling, transformer-BiLSTM combination, domain adaptation, and cross-corpus transfer learning. Further work could involve improving the explanations for readability feedback and prediction for each learner.

References

[1] Jinshan Zeng et al., "InterpretARA: Enhancing Hybrid Automatic Readability Assessment with Linguistic Feature Interpreter and Contrastive Learning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 17, pp. 19497-19505, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[2] Jinshan Zeng et al., "Enhancing Automatic Readability Assessment with Pre-Training and Soft Labels for Ordinal Regression," *Findings of the Association for Computational Linguistics*, pp. 4557-4568, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[3] Zhenzhen Li, Han Ding, and Shaohong Zhang, "Cross-Corpus Readability Compatibility Assessment for English Texts," *IEEE Access*, vol. 11, pp. 101985-101997, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[4] Fengkai Liu, Tan Jin, and John S. Y. Lee, "Automatic Readability Assessment for Sentences: Neural, Hybrid and Large Language Models," *Language Resources and Evaluation*, vol. 59, no. 3, pp. 2265-2296, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[5] Jinshan Zeng et al., "Self-supervised Collaborative Information Bottleneck for Text Readability Assessment," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 24, pp. 25814-25822, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[6] Jinshan Zeng et al., "LearnARA: Automatic Readability Assessment with Deep Linguistic Representation Learner and Contrastive Information Bottleneck," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 34, pp. 992-1003, 2026. [CrossRef] [Google Scholar] [Publisher Link]

[7] Xinying Qiu et al., "Learning Syntactic Dense Embedding with Correlation Graph for Automatic Readability Assessment," *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pp. 3013-3025, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[8] Özkan Aslan, Caner Balım, and Naim Karasekreter, "Advancing Turkish Readability Assessment with Multi-Layer Linguistic Features and Ordinal-Aware Deep Learning," *2025 International Conference on Intelligent Systems: Theories and Applications (SITA)*, Rabat, Morocco, pp. 1-6, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[9] Fengkai Liu, and John Lee, "Hybrid Models for Sentence Readability Assessment," *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 448-454, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[10] Ho Hung Lim, and John Lee, "Improving Readability Assessment with Ordinal Log-Loss," *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications, Association for Computational Linguistics*, Mexico, pp. 343-350, 2024. [Google Scholar] [Publisher Link]

[11] Ahmet Yavuz Uluslu, and Gerold Schneider, "Exploring Linguistic Features for Turkish Text Readability," *arXiv preprint*, pp. 1-10, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[12] Sepideh Ravanbakhsh, and Mohammad Mahmoodi Varnamkhast, "Persian Text Readability Assessment with Hierarchical Transformer-Based Classification Models," *Scientific Reports*, vol. 16, no. 1, pp. 1-11, 2026. [CrossRef] [Google Scholar] [Publisher Link]

[13] Khalid N. Elmadani, Nizar Habash, and Hanada Taha-Thomure, "A Large and Balanced Corpus for Fine-Grained Arabic Readability Assessment," *Association for Computational Linguistics*, pp. 16376-16400, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[14] Justin Lee, and Sowmya Vajjala, "A Neural Pairwise Ranking Model for Readability Assessment," *Findings of the Association for Computational Linguistics*, pp. 3802-3813, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[15] Bruce W. Lee, Yoo Sung Jang, and Jason Lee, "Pushing on Text Readability Assessment: A Transformer Meets Handcrafted Linguistic Features," *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 10669-10686, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[16] Li Zhang et al., "Automatic Text Readability Assessment for Educational Content Based on Graph Representation Learning," *Scientific Reports*, vol. 16, pp. 1-15, 2026. [CrossRef] [Google Scholar] [Publisher Link]

[17] Suna-Şeyma Uçar et al., "Exploring Automatic Readability Assessment for Science Documents within a Multilingual Educational Context," *International Journal of Artificial Intelligence in Education*, vol. 34, no. 4, pp. 1417-1459, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[18] Wenbiao Li, Wang Ziyang, and Yunfang Wu, "A Unified Neural Network Model for Readability Assessment with Feature Projection and Length-Balanced Loss," *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 7446-7457, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[19] Jun Zhao, Readability Assessment of Chinese Linguistic Texts based on Dependent Syntactic Networks, Applied Mathematics and Nonlinear Sciences, vol. 9, no. 1, pp. 1-16, 2024. [Online]. Available: https://www.researchgate.net/publication/378229938_Readability_Assessment_of_Chinese_Linguistic_Texts_Based_on_Dependent_Syntactic_Networks

[20] Yuchen Wang et al., "Automatically Difficulty Grading Method for English Reading Corpus With Multifeature Embedding Based on a Pretrained Language Model," *IEEE Transactions on Learning Technologies*, vol. 17, pp. 474-484, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[21] Mohamed Amine Ouassil et al., "Enhancing Arabic Text Readability Assessment: A Combined BERT and BiLSTM Approach," *2024 International Conference on Circuit, Systems and Communication (ICCSC)*, Fes, Morocco, pp. 1-7, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[22] Yurui Zheng, Yijun Chen, and Shaohong Zhang, "Hierarchical Ranking Neural Network for Long-Document Readability Assessment," *arXiv preprint*, pp. 1-18, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[23] Diego Palma, and Christian Soto, "Combining Large Language Models with Linguistic Features for the Readability Complexity Assessment of Texts," *AHFE International*, vol. 199, 2025. [CrossRef] [Publisher Link]

[24] Catarina Belem et al., "Readability Reconsidered: A Cross-Dataset Analysis of Reference-Free Metrics," *Proceedings of the Fourth Workshop on Text Simplification, Accessibility and Readability*, *Association for Computational Linguistics*, Suzhou, China, pp. 47-69, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[25] Zarah Weiss, and Detmar Meurers, "Assessing Sentence Readability for German Language Learners with Broad Linguistic Modelling or Readability Formulas: When do Linguistic Insights Make a Difference?," *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications*, *Association for Computational Linguistics*, pp. 141-153, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[26] Sowmya Vajjala, "Trends, Limitations and Open Challenges in Automatic Readability Assessment Research," *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, *European Language Resources Association*, Marseille, France, pp. 5366-5377, 2022. [Google Scholar] [Publisher Link]

[27] Ziyang Wang et al., "FPT: Feature Prompt Tuning for Few-Shot Readability Assessment," *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, *Association for Computational Linguistics*, Mexico, pp. 280-295, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[28] Juan Liberato et al., "Strategies for Arabic Readability Modelling," *Proceedings of the Second Arabic Natural Language Processing Conference*, *Association for Computational Linguistics*, Bangkok, Thailand, pp. 55-66, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[29] Mykola Trokhymovych, Indira Sen, and Martin Gerlach, "An Open Multilingual System for Scoring Readability of Wikipedia," *arXiv preprint*, pp. 1-16, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[30] Nizar Habash et al., "Guidelines for Fine-Grained Sentence-Level Arabic Readability Annotation," *Proceedings of the 19th Linguistic Annotation Workshop*, *Association for Computational Linguistics*, Vienna, Austria, pp. 359-376, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[31] Wenjing Pan et al., "Textual Form Features for Text Readability Assessment," *Natural Language Processing*, vol. 31, no. 3, pp. 800-841, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[32] Eugenio Ribeiro, Nuno Mamede, and Jorge Baptista "Automatic Assessment of Text-Complexity Levels in European Portuguese," *Linguamática*, vol. 16, no. 2, pp. 115-139, 2024. [CrossRef] [Publisher Link]

[33] Xiaopeng Zhang, and Xiaofei Lu, "Aligning Linguistic Complexity with the Difficulty of English Texts for L2 Learners Based on CEFR Levels," *Studies in Second Language Acquisition*, vol. 47, no. 5, pp. 1407-1434, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[34] Abdul Aziz et al., "Leveraging Contextual Representations with a BiLSTM-based Regressor for Lexical Complexity Prediction," *Natural Language Processing Journal*, vol. 5, pp. 1-16, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[35] D.A. Morozov, A.V. Glazkov, and B.L. Iomdin, "Text Complexity and Linguistic Features: Their Correlation in English and Russian," *Russian Journal of Linguistics*, vol. 26, no. 2, pp. 426-448, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[36] Yuki Arase, Satoru Uchida, and Tomoyuki Kajiwara, "CEFR-based Sentence-Difficulty Annotation and Assessment," *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, *Association for Computational Linguistics*, pp. 6206-6219, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[37] Nour Rabih, "Noor at BAREC Shared Task 2025: A Hybrid Transformer-Feature Architecture for Sentence-level Readability Assessment," *Proceedings of The Third Arabic Natural Language Processing Conference*: *Shared Tasks*, *Association for Computational Linguistics*, Suzhou, China, pp. 331-342, 2025. [Google Scholar] [Publisher Link]

[38] Mutaz Ayesh, "PalNLP at BAREC Shared Task 2025: Predicting Arabic Readability Using Ordinal Regression and K-Fold Ensemble Learning," *Proceedings of The Third Arabic Natural Language Processing Conference*, pp. 343-349, 2025. [Google Scholar] [Publisher Link]

[39] Sandra Aluisio et al., "Readability Assessment for Text Simplification," *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, Association for Computational Linguistics, Los Angeles, California, pp. 1-9, 2010. [Google Scholar] [Publisher Link]

[40] Rodrigo Wilkens et al., "Exploring Hybrid Approaches to Readability: Experiments on the Complementarity between Linguistic Features and Transformers," *Findings of the Association for Computational Linguistics*, pp. 2316-2331, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[41] Varun Sai Alaparthi et al., "Rating Ease of Readability Using Transformers," *2022 14th International Conference on Computer and Automation Engineering (ICCAE)*, Brisbane, Australia, pp. 117-121, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[42] Susmoy Chakraborty, Mir Tafseer Nayeem, and Wasi Uddin Ahmad, "Simple or Complex? Learning to Predict the Readability of Bengali Texts," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 14, pp. 12621-12629, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[43] Zijie Zeng, Dragan Gasevic, and Guangliang Chen, "On the Effectiveness of Curriculum Learning in Educational Text Scoring," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 12, pp. 14602-14610, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[44] Herianah et al., "Automated Assessment of Text Complexity through the Fusion of AutoML and Psycholinguistic Models," *Forum for Linguistic Studies*, vol. 7, no. 3, pp. 46-62, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[45] XiuHua Zhao, "A Hybrid Deep-Learning and Fuzzy-Logic Framework for Feature-Based Evaluation of English-Language Learners," *Scientific Reports*, vol. 15, no. 14, pp. 1-40, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[46] Muhammad Zulqarnain, and Muhammad Saqlain, "Text Readability Evaluation in Higher Education using CNNs," *Journal of Industrial Intelligence*, vol. 1, no. 3, pp. 184-193, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[47] Ahmad Jaber Mayahi, and Emman Naser Alshatti, "Assessing English-Language Writing and Readability Skills using a Long Short-Term Memory Model," *2023 Computer Applications & Technological Solutions (CATS)*, Mubarak Al-Abdullah, Kuwait, pp. 1-4, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[48] Parahonco Alexandr, and Parahonco Liudmila, "Feature-Level Decomposition of Text Complexity: Cross-Domain Empirical Evidence," *Computer Science Journal of Moldova*, vol. 33, no. 2, pp. 257-280, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[49] G. Pavani et al., "An Interpretable Dual-Level Feedback Approach for Improving Graded Language Simplification and Readability," *International Journal of Advanced Computer Science and Applications*, vol. 16, no. 11, pp. 1-14, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[50] Yi Zuo, "Automatic Generation of ESL Learning Materials based on CEFR Levels Using Reinforcement-Tuned Large Language Models," *Discover Artificial Intelligence*, vol. 6, no. 1, pp. 1-24, 2026. [CrossRef] [Google Scholar] [Publisher Link]