A Dual-Path Gated Attention-Based Deep Learning Model for Automated Essay Scoring Using Linguistic Features

Qin Jie[1]*, Congling Huang[2]
Shaanxi Fashion Engineering University, Xi'an, Shaanxi, 712046, China[1]
Shaanxi Vocational and Technical University of Agriculture and Forestry, Shaanxi, Yangling 712000, China[2]

***Abstract*—Automated Essay Scoring (AES) has become a critical tool for scaling writing assessment in modern education. However, existing AES models often struggle to effectively evaluate both the syntactic structure and semantic meaning of essays while maintaining interpretability and fairness. This study presents a novel deep learning-based model that integrates syntactic and semantic analysis using an improved LSTM architecture. The model employs a dual-path structure: one path processes semantic representations using BERT-tokenized input, while the other captures syntactic patterns via part-of-speech sequences. These paths are fused using a gated mechanism and enhanced through multi-head attention to emphasize important linguistic cues. Additional student metadata, such as grade level and gender, is also incorporated to improve personalization and fairness. The model jointly predicts both holistic and grammar scores, trained and evaluated on the ASAP 2.0 dataset. Performance is measured using multiple statistical metrics, including MAE, MSE, RMSE, R^2, Pearson's r, and Spearman's ρ. The proposed model achieves a high prediction accuracy of 92%, significantly outperforming traditional and single-path models. These results demonstrate the model's ability to capture both surface-level and deep linguistic features, offering a robust, interpretable, and scalable solution for automated writing evaluation.**

***Keywords*—Attention mechanism; deep learning; essay scoring; gated fusion; linguistic features; semantic encoding; syntactic representation**

I. Introduction

Automated Essay Scoring (AES) has emerged as a promising solution to meet the growing demands of large-scale writing assessment in education. It offers scalability in evaluation. It ensures objectivity and consistency in assessing student writing [1]. Traditional manual scoring is thorough, however, it is often time-consuming [2]. It is also costly and prone to inter-rater variability [3]. AES systems aim to provide fast feedback. They also aim to provide reliable evaluation. These systems maintain fairness during scoring [4]. They minimize human bias in evaluation. The growing role of digital learning platforms has increased the demand for AES. The increased reliance on online assessments further emphasizes the importance of developing efficient AES systems. It also highlights the need for accurate AES systems.

Over the years, AES has evolved through multiple stages. It began with rule-based systems. These systems used handcrafted feature extraction. Later, statistical and machine learning models were introduced. Currently, research is shifting towards end-to-end deep learning solutions [5]. Early AES systems heavily relied on surface-level linguistic features. These features included grammar error counts. They also included sentence length and vocabulary usage. Additional features were part-of-speech distributions and syntactic complexity. These systems demonstrated moderate performance. However, they struggled to capture deeper semantic relationships. They could not fully understand contextual meaning. They failed to capture discourse coherence. They also lacked the ability to model structural flow within essays. Furthermore, these systems often lack adaptability. They did not generalize well across different topics and domains [6]. This was due to the handcrafted nature of their features.

In response to these limitations, recent research has explored the potential of deep learning models [7]. These include Recurrent Neural Networks (RNNs). They also include Long Short-Term Memory (LSTM) networks. Convolutional Neural Networks (CNNs) have also been used. Transformer-based models such as BERT and RoBERTa have gained popularity [8]. These models have demonstrated superior capability in feature learning. They can automatically extract syntactic representations from text. They can also capture semantic relationships from raw data. Some studies have extended these deep models. They have incorporated hierarchical structures. They have also used multi-task learning. Hybrid feature fusion techniques have been explored. Attention mechanisms have been added to improve focus on relevant parts of the essay. Despite these advancements, two major challenges still persist in AES research [9]. First, many models cannot capture syntactic structure and semantic meaning at the same time. They fail to unify both aspects in a single framework. Second, most models are not interpretable. They also lack the ability to generalize across different essay prompts. They struggle to perform consistently across diverse student demographics.

To address these gaps, researchers have proposed hybrid models [10]. These models combine neural embeddings with manually engineered linguistic features. Their goal is to leverage the strengths of both approaches. Neural models offer automated feature extraction. Linguistic features provide human-readable grammar and readability metrics. This combination enhances scoring accuracy. It also improves transparency of predictions. However, these hybrid models

*Corresponding Author.

introduce added complexity. They are harder to design and train. Many of them still lack effective coherence modeling. They struggle to capture relationships across multiple levels of essay structure [11]. Other recent works have shifted to Transformer models. Large Language Models (LLMs) have also been explored [12]. These models show strong scoring capabilities. However, they are data-hungry. They require large amounts of training data. They are computationally expensive to run. They are also hard to interpret. As a result, there is a growing need for AES models that can strike a balance. The ideal model should be interpretable. It should be computationally efficient. It must model essay coherence. It should also deliver high scoring accuracy.

In this paper we show an enhanced LSTM architecture based deep LSTM model. The model is intended only for syntactic and semantic assessment of English essays. Our method presents a dual path gated model. It feeds essays through two parallel but complementary channels. The first direction is the one focusing on semantics. It takes as input tokenized text which is pre-processed with a BERT vocabulary. A second approach is a syntactic one. It employs POS-sequences from SpaCy. The second layer is for every path through a bi-LSTM layer. This is followed by a multi-head attention mechanism. This is how it helps the model to learn important patterns and interactions. A sentence level gating mechanism is employed. It builds a dynamic connection between the two roads. This gives the model the freedom to choose when to take significance from syntax. It also learns to pay more attention to semantic features. Apart from such a mirror-path architecture, we also incorporate student metadata into the model. Features in the metadata are the grade level and the gender. Personalization is facilitated with this knowledge. It further encourages a more level playing field with scoring, with formulated research questions are.

Research Question 1: How can a deep learning model effectively integrate both syntactic and semantic features to enhance automated essay scoring (AES)?

Research Question 2: Does the integration of student metadata and dual linguistic paths improve the fairness and personalization of AES systems?

Research Question 3: Can the proposed SYNSENNET model outperform traditional and single-path AES models in terms of prediction accuracy and interpretability?

The last model is multiple outputs. It estimates holistic essay scores and grammar scores in a joint manner. We built and tested this model with the ASAP 2.0 dataset. It is a publicly available real-world dataset which has been widely utilized. It includes student essays that have been annotated with more than one scoring attribute. We adopted conventional regression performance measures to evaluate the models. This measure involves Mean Absolute Error (MAE). They also contain Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). The Coefficient of Determination (R^2) was determined. We also fad-ea12301|Discussion e12301fused the Pearson Correlation Coefficient (r). Finally, we evaluated Spearman's Rank Correlation Coefficient (ρ). Our model showed a high predictive value of 92%. This finding indicates high consistency with human annotations. It is shown that the model's performance is good enough in holistic and grammar judgments.

Key contributions of this study include:

- Development of a novel dual-path deep learning architecture that jointly models syntactic structure and semantic content.
- Integration of multi-head attention and gating mechanisms for dynamic fusion of linguistic features.
- Inclusion of student metadata to improve scoring fairness and personalization.
- Simultaneous prediction of both holistic and grammar scores in a multi-output framework.
- Empirical validation on the ASAP 2.0 dataset, showing significant performance gains over traditional and single-path models.
- Application of diverse evaluation metrics to offer a robust and interpretable assessment of model performance.

The remainder of this paper is structured as follows. Section II overviews known techniques for AES and discusses their strengths and weaknesses. The proposed methodology in Section III presents the strategy of the current study in a nutshell, before describing the proposed framework in details, along with its two paths and the important components. The Model Workflow describes the overview of what the model is doing, and the Computational Efficiency explains a comparison of computational complexity. Details about the training setup, prepossessing and implementation can be found in Section IV. The dataset describes the ASAP 2.0 dataset and why this dataset was considered. Performance Evaluation Metric In this section, biostatistical metrics which evaluate our method are introduced. In Section V, we report on the experiments and discuss on the performance with comparison to related work. Lastly, Section VI reviews the key findings and provides future research directions.

II. Related Work

Early AES systems strongly made use of handcrafted linguistic features (such as counts of parts-of-speech, syntactic structure, cohesion markers) in conjunction with traditional ML regression models, summary analysis shown in Table I. In contrast, deep learning methods automatically extract features from text, however, they are often not interpretable. For instance, Kumar and Boulanger [13] used multi-layer perceptron's to learn from 1592 linguistic indices (which can be used to assess improved text cohesion, lexical diversity/sophistication, syntactic complexity, and more) and found that deeper networks increased score prediction performance by around 10%. They also used SHAP for feature-attribution and to identify which linguistic features affect these scores. Similarly, Kumar and Boulanger [14] created an MLP that predicts rubric-based sub-scores and subsequently the overall essay score; their AES achieves a high holistic QWK of 0.78 (exceeding human agreement). These results show how large selections of linguistic features combined with deep networks can offer very strong accuracy

and provide some degree of interpretability of the scoring justification.

Recent research was also directed to hybrid models, which combine linguistic features with neural embeddings. Uto et al. [15] concatenate 25 essay-level features (e.g., length, vocabulary, syntax) to an intermediate essay representation in a DNN (LSTM or BERT) and report substantial QWK gains across all prompts. Cho et al. [10] hybridize RoBERTa sentence embeddings-Word embeddings with manually crafted grammar/readability features by an XGBoost regressor, to obtain a QWK of 0.941 on the Kaggle ASAP dataset as well. Doi et al. [16] utilize grammatical knowledge through multi-task learning: predicting holistic and sub-scores for grammar and weighting error features by difficulty (IRT); the addition of grammar features and MTL led to large gains in performance. Such hybrid approaches tap into complementary information by using dual-path inputs (neural embeddings + Ling. feat.) and are more effective in terms of accuracy and interpretability.

Transformer and BERT-inspired models have gained popularity in AES. Ludwig et al. [17] applied transformer encoders (pre-trained on essays) and observed that they did significantly better than a baseline bag-of-words logistic model on a student email classification task. Xue et al. [18] presented a hierarchical BERT model with multi-task fine-tuning towards multi-trait scoring; they modeled essays in different granularities and weighted segments using attention, achieving +4.5% (ASAP dataset) and +8.1% (Chinese EFL essays) QWK gain over baseline BERT. Wang et al. [19] proposed a multi-scale essay representation in BERT, learning token-, segment-, and document-level embeddings simultaneously, with transfer learning, which obtain near state-of-the-art results on ASAP and generalize well to a different Commonly readability dataset. These works demonstrate that hierarchical and multi-scale BERT-based models can capture both fine-grained and global features of essays for better scoring in long essays. LLMs are currently under evaluation for usage in AES. for traditional AES systems (e.g. rule-based Jess/Writer) using as our test data for these systems 1,400 essays by Japanese L2 writers, we compared GPT-4, a fine-tuned BERT, and a Japanese-specific LLM with the work of Li and Liu [20] on the same set of essays. They discovered that GPT-4 also performs the best in holistic accuracy, more even than BERT and previous AES tools, and products writer proficiency better than BERT. Quah et al. [21] considered ChatGPT-4 as an essay grader for 300 essays of dental school: GPT's scores strongly correlated with human graders ($r \approx 0.83$) and presented a very good inter-rater reliability, thus suggesting that LLM are reliable AES agents. Atkinson and Palma [22] build on this to propose an LLM-based hybrid model that combines lexical/discourse features with neural context from a large model; they show that this ensemble outperforms both shallow-feature and pure neural baselines on standard essay benchmarks. These initial findings suggest instruction-tuned LLMs (GPT4) can rival or surpass legacy AES methods, particularly when combined with simple prompt engineering and feature fusion.

A second concern is the generalization over prompts. Wang et al. [23] cast AES as a meta-learning task: the meta-learner leverages multiple source prompts to guide model to distributions of unseen target prompts and improves cross-prompt accuracy on ASAP. Jiang et al. [24] deal with domain (prompt) generalization by learning disentangled representations for essays: they decompose prompts-invariant and prompts-specific features via contrast-centric and counterfactual training, which in turn leads to better performance on unseen prompts on ASAP and TOEFL. Such work suggests that separating content (essay meaning) from prompt context can enhance the AEs robustness when it is tested in new essay topics.

However, despite the strong predictors, deep AES frameworks are often uninterpretable. Studies of AES (Misgna et al., [7]) emphasize that state-of the- art model succeed at capturing complex pat- terns without indicating which features lead to scores. So explainable AES is still an area of focus: e.g. Kumar and Boulanger claim that predicting rubric sub-scores helps interpretability, and others use feature attribution (e.g. SHAP) to connect neural predictions to linguistic characteristics. In summary, recent AES research is characterized by a wide shift that ranges from traditional feature-based approaches to complex deep/LLM models, where hybrid and hierarchical architectures that combine linguistic knowledge with high scoring ability are recently favored. The study by Beseiso et al. [8] focuses on developing an advanced automated essay scoring (AES) system to meet the growing demands of e-learning and higher education. Traditional essay scoring methods often fail to capture the coherence and deeper structural elements of essays. To address this, the researchers propose a transformer-based neural network that combines RoBERTa, a powerful pre-trained language model, with a Bi-directional Long Short-Term Memory (Bi-LSTM) layer. The core idea is to enhance RoBERTa's contextual language understanding with Bi-LSTM's ability to model sequential dependencies, thereby overcoming the document length limitations associated with transformer models.

This integrated model is also more effective at modeling long-form essays, since it maintains coherency and long-range dependency between text. The problem is posed as a regression and is tested in Kaggle's ASAP dataset. Experimental results demonstrate that the proposed model achieves better performance than classical NLP pipelines, deep learning models, and combined architectures.

It shows better alignment with human raters, especially in the writing of essays, and it's proved to be a promising tool to deliver for a computerized grading system at universities. However, the model is not equipped with explicit syntactic modeling and deep semantic reasoning. Although it learns semantic relationships, it does not consider structural aspects, such as grammatical rules and dependency relations, as black box constraints. The latter, however, may not fully represent fine-grained linguistic features that are required for deep syntactic and semantic analysis of English essays.

TABLE I. SUMMARY OF RECENT AES MODELS, THEIR FEATURES, PERFORMANCE, AND LIMITATIONS

Ref	Model Name	Dataset	Feature	Result (Acc)	Limitation	Strength	Area
Kumar & Boulanger (2020)	MLP with SHAP	Custom (1592 feats)	1592 handcrafted linguistic features	+10% improvement	Black-box nature; interpretability via SHAP only	High accuracy; interpretable via SHAP	Deep + Feature-Based AES
Kumar & Boulanger (2021)	MLP for sub-scores	Not Specified	Rubric-based sub-score prediction	QWK = 0.78	Requires labeled sub-scores	Exceeds human agreement; interpretable scoring	Deep + Interpretable AES
Uto et al. (2020)	Hybrid DNN	ASAP	25 handcrafted essay-level features + LSTM/BERT embeddings	Substantial QWK gains	Feature engineering needed	Hybrid of DNN + features; effective accuracy	Hybrid AES
Cho et al. (2024)	Hybrid RoBERTa + XGBoost	ASAP	RoBERTa + Word Embeddings + handcrafted grammar/readability features	QWK = 0.941	Complexity in pipeline	Very high accuracy	Hybrid AES
Doi et al. (2024)	Multi-Task Learning	Not Specified	Grammar features + difficulty-weighted error features (IRT)	Large performance gains	IRT modeling required	Exploits grammar deeply; effective in MTL setup	Hybrid/MTL AES
Ludwig et al. (2021)	Transformer Encoder	Student emails	Pretrained transformer on essays	Better than baseline	Domain-specific; task not essay scoring directly	Strong generalization; simple model	Transformer-based AES
Xue et al. (2021)	Hierarchical BERT	ASAP, Chinese EFL	Hierarchical structure + attention weighting on segments	+4.5% (ASAP), +8.1% (EFL)	Training complexity	Granular modeling of traits; multi-task fine-tuning	Hierarchical BERT AES
Wang et al. (2022)	Multi-scale BERT	ASAP, CommonLit	Token/Segment/Document-level embeddings + Transfer Learning	Near SoTA	High resource requirement	Strong generalization to unseen data; captures both local/global info	Multi-scale BERT AES
Li and Liu (2024)	GPT-4 vs. BERT vs. legacy AES	Japanese L2 essays	GPT-4 vs. BERT vs. legacy AES	GPT-4 > BERT & others	Limited prompt control	GPT-4 best in holistic accuracy; handles proficiency grading well	LLM AES
Quah et al. (2024)	ChatGPT-4 as grader	Dental Essays	ChatGPT-4 as grader	r ≈ 0.83 (w/ humans)	Prompt sensitivity	Strong inter-rater reliability; high human correlation	LLM AES
Atkinson & Palma (2024)	Hybrid LLM	Essay Benchmarks	LLM + Lexical/Discourse Features	Outperforms baselines	Requires feature fusion	Strong hybrid model with high interpretability and accuracy	LLM + Feature AES
Wang et al. (2025)	Meta-learning AES	ASAP	Meta-learner across prompts	Improved cross-prompt	Complexity in training across prompts	High generalization to unseen prompts	Prompt-Generalization AES
Jiang et al. (2023)	Disentangled AES	ASAP, TOEFL	Prompt-invariant + Prompt-specific features via contrastive learning	Better on unseen prompts	Training complexity; separation logic	Robust to unseen prompts; separates content and prompt	Prompt-Generalization AES
Misgna et al. (2024)	Explainable AES	Not Specified	Focus on interpretability	Not quantified	No clear link between features and scores	Highlights ongoing need for interpretability in deep AES	Explainable AES
Beseiso et al. (2024)	RoBERTa + BiLSTM	ASAP	RoBERTa + BiLSTM for long-form coherence modeling	Beats hybrid & deep	Lacks explicit syntactic/deep semantic modeling	Excellent coherence capture; scalable for long essays	Transformer Hybrid AES

A. Limitations of Existing Studies

Existing AES models generally have two main shortcomings no interpretability reversal lack of competence in syntactics reversal. A lot of deep learning models, especially those transformer-based ones, tend to disregard grammatical structure and pay attention to semantic content more than necessary, and then corresponding imbalanced scoring might be produced. Moreover, models that use heuristically derived features or process taken sequences in a single path fail to generalize between different writing styles and student populations. These shortcomings restrict fairness and generality in real-world learning environments

III. PROPOSED RESEARCH METHODOLOGY

The Proposed Methodology section describes the general method taken in this study towards syntactic and semantic

essay assessment. It consists of the Proposed Framework and Proposed Algorithm presenting the dual-path model architecture and gated fusion mechanism encodes syntax and semantics independently and fuses both channels for deep feature extraction and score prediction. Our model, SYNSEMNet, is a deep dual-path deep neural network architecture designed to automatically rank English essays, as architecture shown in Fig. 1. It rates two dimensions of writing: quality and grammar. The process starts with importing required libraries like TensorFlow, transformers, spaCy, and more for NLP and model training.

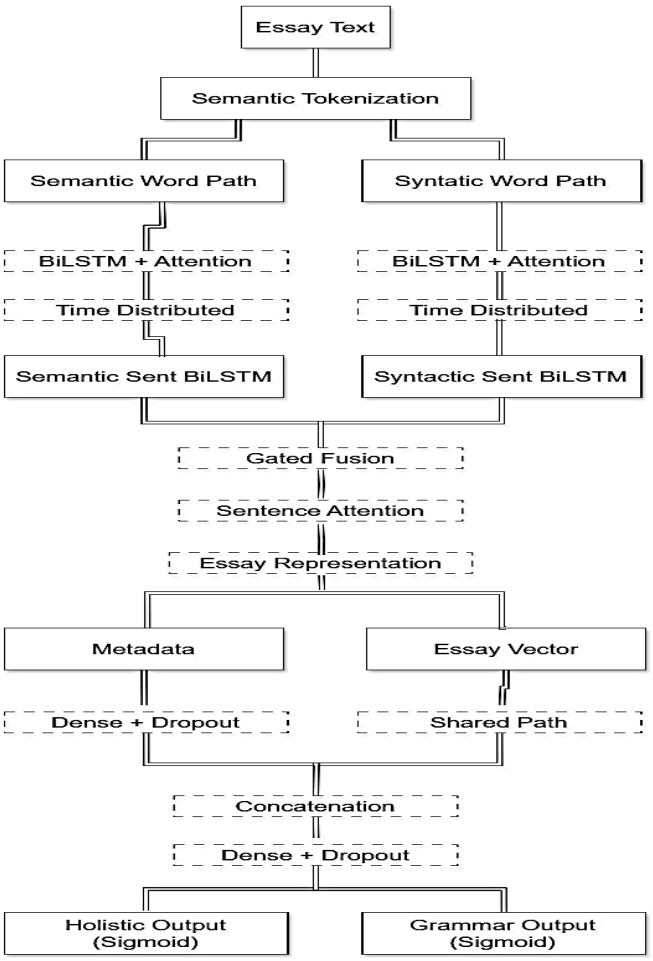

Fig. 1. Block and flow diagram of SYNSEMNet architecture.

The dataset employed is the ASAP 2.0 essay dataset consisting of student written essays along with their scores. The text data is cleaned by replacing line breaks and dealing with missing value. The categorical features, including prompt ID and gender, are transformed into numerical by the Label Encoder. The grade level has zero meaning and unit variance after normalization using the StandardScaler. Target scores (holistic and grammar) are scaled to the range of 0 to 1. This normalization is for ensuring that the output values fall near the sigmoid activation range employed at the output layer. Each essay is preprocessed to obtain both syntax and semantics. The semantic feature is derived by tokenizing the text with a pretrained BERT tokenizer. This behavior allows to maintain the meaning of words. For the extraction of syntactic features, we make use of spaCy, which provides part-of-speech (POS) tags per word. Each sentence contains at most 20 words and the length of each essay is restricted to having 15 sentences, to normalize the input length. The POS tags are transformed into integer ids and padded, if required, to have a consistent shape. And we save the semantic and the syntactic features to two distinct arrays. In addition, related metadata (e.g., encoded gender and scaled grade level) are maintained independently. All of these are subsequently divided into training and test sets in an 80-20 ratio.

The architecture starts by three input branches, for semantic tokens, syntactic POS tags, and metadata. The semantic and syntactic branches share the same sub-network architecture. Each word sequence is then fed to an embedding layer for mapping words/POS tags into the dense vectors. Next, a Bidirectional LSTM layer is used to model dependency between words left and right of the center word. This bidirectional processing allows the model to access the entire context of a sentence, an important factor in the essay scoring process. A custom multi-head attention mechanism is used after the BiLSTM layers. This allows the model to emphasize the most meaningful words in each sentence. Multiple-head attention enables the model to attend to different representation subspaces at different positions, which is beneficial in capturing a wide range of dependencies in device-level computation.

Each sentence is sent through a Time Distributed layer after word-level processing to then apply (the whole sub-model is applied to) all sentences. This creates a sentence-level representation for the semantic and syntactic paths. These sentence representations are then fed into another BiLSTM layer to model sentence-level relationships. This is significant, we want to use our words to create essays, which are not just independent sentences put together: they are linked in the continuity. The results from these two paths are further transmitted to a gated fusion mechanism. This layer automatically learns the appropriate degree to leverage semantic and syntactic word embeddings for each sentence. Finally, a soft gate is computed with a sigmoid activation. This gate modulates the joint semantic and syntactic outputs. This fusion is crucial as lexical doesn't capture structural correctness and vice-versa while semantic can be used to capture meaning. Then a sentence-level attention is performed. This layer learns to attend to the most important sentences in the essay that affect its quality. Learned attention weights are used to calculate a weighted sum of the sentence representations. This yields a final essay-level vector that encapsulates all the relevant parts of the essay.

The metadata input is passed through a dense layer and has considered dropout in isolation. This includes context information about the student (e.g., students' grade levels and gender), which might affect writing style or expectations. The weighted sum from the attention mechanism, h at $\in$ R d, and the metadata vector, q, are concatenated to form a joint representation. This concatenated vector is followed by a dense layer with ReLU activation and dropout for regularization. Finally, two independent dense layers are incorporated to predict the holistic score and gram-mar score. Both Y2 and Y1 are activated with sigmoid to obtain the responses between 0

and 1. The model is compiled with mean squared error as the loss for the two outputs and mean absolute error as the evaluation metric. We use Adam as the optimizer with learning rate = 0.0001, as it is good for training deep networks. The model is trained for 6 epochs with a batch size 16. A subset of the training data is used for validation in the training. The trained model is then tested on the test set. It outputs predicted holistic and grammar scores, and the prediction results are evaluated via measurement methods such as MAE, MSE, RMSE, R^2, Pearson correlation, and Spearman correlation. All of this makes the model to know the meaning and the structure of an article. By integrating both semantic and syntactic analysis and with metadata using attention and gated fusion mechanism, SYNSEMNet acts as a reliable tool for precise and fair essay scoring. Fig. 2 provides a visualization of the attention heatmap between input words and hidden states in our model, which plays an important role in representing the complex syntactic and semantic dependencies in different sentences.

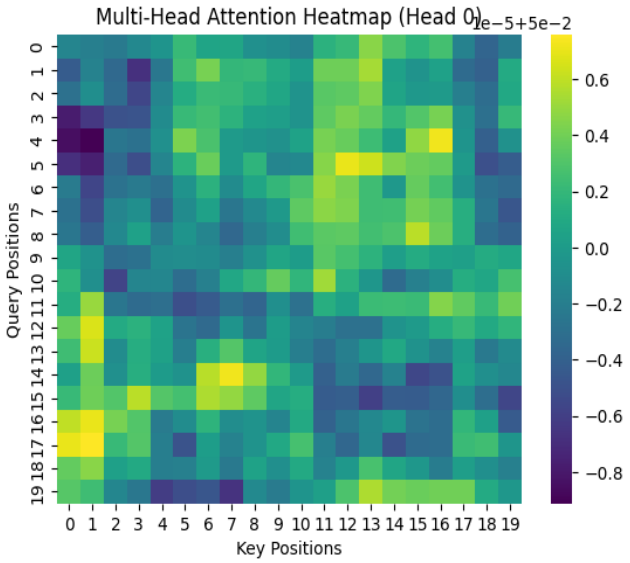

Fig. 2. Multi-head attention heatmap showing word-level focus patterns in SYNSEMNet's semantic encoding.

Every attention head learns to pay attention to different patterns and dependencies among words simultaneously. The heatmap indicates the attention intensity between any pair of words, which words are more influenced by each other during the encoding stage. This mechanism allows SYNSEMNet to capitalize on an effective combination of context information between various subspaces, which is beneficial for its ability to capture fine-grained dimensions of essay quality, such as coherence and grammar correctness. Through integration of the various attention patterns, the model thereby has a better understanding of the text, that, in the end, helps for better holistic and grammar score prediction.

A. Proposed Algorithm

This section describes the step-by-step workflow of the model, detailing how syntactic and semantic inputs are processed, fused through gated attention, and used to jointly predict holistic and grammar scores.

B. Justification of Model Selection

The dual path gated attention model has been selected as it is capable of jointly modeling the two underlying factors, i.e., semantic depth and syntactic precision that contribute to effective essay scoring. Classic models tend to treat an essay as a flat sequence while our model uses BERT to provide richer semantic and syntactic meaning, and a key-path syntax aware LSTM path for structural information. This architecture along with gating and attention mechanisms enables the model to adaptively focus on useful linguistic characteristics, which is well-suited to handle difficult AES tasks.

The algorithm begins by taking an essay, represented as text E, along with metadata m, as input. Each essay in the dataset is first divided into sentences, forming a set $S = \{s1, s2, \ldots, sM\}$. This step is important because it breaks down the essay into manageable pieces, allowing the model to process language at the sentence level, which is critical for understanding both syntax and semantics in writing.

For every sentence s_i in the set S, the sentence is tokenized using BERT's tokenizer, converting it into a sequence of token Ids $S_i \in Z^N$, where N is the maximum number of words per sentence. Alongside this, part-of-speech (POS) tags $P_i \in Z^N$ are extracted using the spaCy tool. Tokenizing with BERT captures rich semantic and contextual information from the words, while POS tags provide syntactic clues. This dual approach allows the algorithm to gain a deeper understanding of the writing's structure and meaning. The tokenization is given by Eq. (1)

$$S_i = \text{BERTTokenize}(s_i) \in Z^N \quad (1)$$

After all sentences have been tokenized and POS-tagged, padding is applied to form fixed-size matrices $S_i \in Z^{M\times N}$ and $P_i \in Z^{M\times N}$. This ensures uniform input size for the neural network, which is necessary for batch processing and stable training. Next, embeddings are generated: E_s from BERT embeddings for the tokens, and E_p from POS embeddings for the syntactic tags. Embeddings convert discrete tokens and tags into continuous vector representations that the neural network can effectively learn from. The embeddings are shown in Eq. (2).

Algorithm: Gated BiLSTM with Multi-Head Attention for Deep Syntactic and Semantic Assessment of English Essays

Input: Essay Text Є, Metadata *m*
Output: Predicted Holistic, Grammar Score

1. **For each** essay $E \in$ dataset, **do**
2. $S = \{s_1, s_2, \ldots, s_M\}$ //Tokenize $\mathcal{E}$ into M sentences
3. **For each** sentence $s_i \in S$, **do**
4. $S_i = \text{BERTTokenize}(s_i) \in Z^N$ //Tokenize s_i into N words and convert to token IDs
5. Extract POS tags $P_i \in Z^N$ *using spaCy*
6. ***End For***
7. $Pad = \{S \in Z^{M\times N},\ P \in Z^{M\times N}\}$

8. ***End For***
9. $E_s = \text{Embedding}_{\text{BERT}}(S), \quad E_p = \text{Embedding}_{\text{POS}}(P)$
10. **For each** sentence $i = 1$ to M **do**
11. $H_{s,i} = \text{BiLSTM}(E_{s,i}) \in R^{N\times H}$ $H_{p,i} = \text{BiLSTM}(E_{p,i}) \in R^{N\times H}$ //BiLSTM + Multi-head Attention
12. $A_{s,i} = \text{MHA}(H_{s,i}), \quad A_{p,i} = \text{MHA}(H_{p,i})$ //Multi-head Self Attention
13. ***End For***
14. $Stack = \{A_s \in R^{M\times N\times H}, \quad A_p \in R^{M\times N\times H}\}$
15. $S_s = \text{BiLSTM}_{\text{sent}}(A_s) \in R^{M\times H}$ $S_p = l\text{BiLSTM}_{\text{sent}}(A_p) \in R^{M\times H}$ //Bi-LSTM across sentence
16. $G = \sigma\backslash big(S_s \cdot W_g + b_{g\backslash big})$ //Compute gate
17. $F = G \odot S_s + (1 - G) \odot S_p$ //Fuse features
18. $\text{m}' = \text{Dropout}\backslash\text{big}(\text{ReLU}(\text{m} \cdot \text{W}_\text{m} + \text{b}_\text{m})\backslash\text{big})$ //Meta Data Fusion
19. $z = [v \parallel m'], \quad h = \text{Dropout}\backslash big(\text{ReLU}(z \cdot W_z + b_z)\backslash big)$
20. $\widehat{y_{hol}} = \sigma(h \cdot W_{hol} + b_{hol})$
21. $\widehat{y_{gram}} = \sigma(h \cdot W_{gram} + b_{gram})$ //Output

$$E_s = \text{Embedding}_{\text{BERT}}(S), \quad E_p = \text{Embedding}_{\text{POS}}(P) \quad (2)$$

For each sentence index i, the model applies a Bi-directional LSTM (BiLSTM) separately on the semantic embeddings $E_{s,i}$ and the POS embeddings $E_{p,i}$, producing hidden states $H_{s,i} \in R^{N\times H}$ and $H_{p,i} \in R^{N\times H}$, where H is the number of LSTM units. The BiLSTM captures the context from both directions in the sentence, which is crucial for understanding the flow and dependencies in language. Then, multi-head attention (MHA) is applied to these hidden states, producing attention-weighted representations $A_{s,i}$ and $A_{s,i}$. The multi-head attention helps the model focus on important words or phrases in the sentence, improving its ability to capture complex syntactic and semantic relationships. The multi-headed attention is given by Eq. (3)

$$A_{s,i} = \text{MHA}(H_{s,i}), \quad A_{p,i} = \text{MHA}(H_{p,i}) \quad (3)$$

Once all sentences are processed, the attention outputs A_s and A_p are stacked into tensors with shapes $M \times N \times H$. Another BiLSTM layer runs at the sentence level across these stacked tensors, yielding sentence-level features $S_s \in R^{M\times H}$ and $S_p \in R^{M\times H}$. This step captures interactions between sentences and how they contribute to the overall essay structure. The sentence level feature extraction is given by Eq. (4) and Eq. (5).

$$S_s = \text{BiLSTM}_{\text{sent}}(A_s) \in R^{M\times H} \quad (4)$$

$$S_p = l\text{BiLSTM}_{\text{sent}}(A_p) \in R^{M\times H} \quad (5)$$

A gating mechanism is then introduced, where a gate G is computed by applying a sigmoid activation on a linear transformation of S_s. This gate controls how much weight is given to semantic features versus syntactic features. The fused feature F is created by combining S_s and S_p using element-wise multiplication with the gate G and its complement $1 - G$. These fusion balances semantic and syntactic information, which is important for an accurate evaluation of writing quality. Calculation of gate value is given by Eq. (6).

$$G = \sigma(S_s \cdot W_g + b_g) \quad (6)$$

The metadata mm undergoes a transformation through a fully connected layer with ReLU activation and dropout, producing m. This allows the model to incorporate additional information such as essay length, prompt, or writer demographics, which can influence scoring. The semantic-syntactic fused vector v is concatenated with this transformed metadata to form z, which then passes through another fully connected layer with ReLU and dropout to create the final hidden representation h.

Finally, two outputs are generated from h by applying sigmoid activations on separate linear layers. These outputs are the predicted holistic score $\widehat{y_{hol}}$ and the predicted grammar score $\widehat{y_{gram}}$. These predictions reflect the overall quality and grammatical correctness of the essay, which are the main targets of the research. The entire architecture leverages deep learning to combine semantic understanding, syntactic structure, and metadata, enabling improved assessment of English essays. This approach, tested on the ASAP2.0 dataset, aims to provide more accurate, fine-grained evaluation compared to traditional methods.

C. Complexity Comparison

The SYNSEMNet model processes each essay by first dividing it into MM sentences and then tokenizing each sentence into N tokens. The main computational steps include tokenization and POS tagging, which scale linearly with the number of sentences and tokens $O(M \times N)$. However, the most computationally expensive parts are the BERT embedding extraction and multi-head attention applied at the sentence level, which have quadratic complexity with respect to the token length of each sentence $O(M \times N^2)$. Additionally, the BiLSTM layers applied both at the token level and the sentence level add complexity proportional to $O(M \times N \times H)$ and $O(M \times H^2)$ respectively, where HH is the hidden size. Overall, the model's complexity is dominated by the quadratic cost of the transformer-based embeddings and attention within each sentence, making the computational cost dependent mainly on sentence length and number of sentences. Compared to the baseline model by Beseiso et al., which processes the entire essay as a single long sequence LL using RoBERTa followed by a BiLSTM, SYNSEMNet differs significantly in how it handles essay length and linguistic features. Beseiso's approach involves a quadratic complexity $O(L^2)$ from the transformer applied on the entire essay token sequence, which can be costly and limited by maximum input length constraints. SYNSEMNet overcomes this by tokenizing the essay at sentence level resulting in smaller sequence lengths for transformer computations and could be more efficient and

scalable, therefore. Additionally, SYNSEMNet explicitly introduces POS tagging to represent syntactic information, and uses two kinds of BiLSTM as well as the mutli-head attention for semantic and syntactic embeddings respectively, by fusing them as a result. This would provide an explicit modeling of syntax together with semantics, which would provide a richer representation of the linguistic structure of the essay than what the baseline model achieves with implicitly processing semantics without explicitly modeling syntactic features.

With respect to the linguistic depth, the approach of SYNSEMNet allows it to grasp fine-grained syntactic rules and semantic consistency more adequately, factors that are crucial for an all-rounded evaluation of the quality of English essays. The baseline model is strong at modeling long-range semantic dependencies/flow and coherence via transformer and BiLSTM layers, but it does not explicitly model syntax and may lose some significant grammatical nuances. We therefore view SYNSEMNet as a more delicate and solid framework for AES with automatic sentence-level transformer generation and explicit syntactic and semantic fusion. Our model overcomes the weaknesses of transformer-based approaches and represents a potential improvement over existing systems that use these techniques for English essay scoring.

IV. Experimental Setup

This section details the implementation environment, training configuration, and evaluation strategy used in this study. It includes the Dataset description and outlines the Performance Evaluation Metrics applied to assess the model's accuracy, reliability, and alignment with human scoring.

A. Dataset

ASAP2.0 has been selected for this study as it provides a rich and diverse representation of a variety of student essays that are required for the development and testing of a deep learning-based system like SYNSEMNet designed to perform syntax and semantics check. Contrary to large datasets, displayed in Table II, ASAP2. 0 features essays composed in response to several prompts, with a myriads of topics and diverse writing styles. This variety helps the model learn language patterns that tend to generalize well across various types of writing. Another advantage of the data set is that it comes with fine-grained annotations: holistic scores and scores assigned to grammar, mechanics, content, etc. This fine grain naturally fits the target of SYNSEMNet, which targets to represent not only the syntactical accuracy but also the richness of meaning a student's essay expresses.

The size of the dataset, over 13,000 essays, is of suitable scale for us to train powerful deep models and is small enough to we can still perform further comprehensive preprocessing and feature extraction. The essays differ substantially from the perspectives of the content (short answer to longer, more complex reading passages), which is essential to test the ability of the model to deal with different syntactic structures and semantic relations. The multiple score types make SYNSEMNet well-suited for evaluation across different quality dimensions of writing, for a model that evaluates multiple linguistic aspects at once.

TABLE II. Descriptive Statistics of the ASAP 2.0 Dataset

Feature	Statistic	Relevance for SYNSEMNet
Total Essays	13,202	Large enough for deep learning training
Number of Prompts	8	Diverse topics enhance generalization
Essay Length (words)	Min: 50	Captures very short writing
	Max: 730	Captures long and complex essays
	Mean: 350	Average essay length for model learning
Vocabulary Size	15,500 unique tokens	Diverse vocabulary for semantic richness
Holistic Score Range	0 to 60	Reflects wide quality levels
Holistic Score Mean	~35	Average writing quality
Holistic Score Std. Dev.	~10	Variation enabling model differentiation
Grammar Score Range	0 to 30	For focused syntactic evaluation
Grammar Errors Present	In ~20% of essays	Supports syntactic error learning
Average Sentence Count	~15 sentences per essay	Reflects syntactic complexity
Score Distribution Skew	Slight positive skew	More essays near lower scores
Missing Data	<1%	Negligible, good data quality

B. Performance Evaluation Metric

These metrics are the following: MAE, MSE, RMSE, R^2, r and ρ. Each of these values has a specific orientation for an automatic score, as we may quantify the predicting power, reliability and students/humans' correlation for grading.

1) Significance of validation and relative assessment: It is of great importance to perform sound validation of AES systems so that they can be trusted and fair. We demonstrate this empirically through thorough experimentation using a multitude of statistical measures and comparison to classical baselines. The addition of Pearson's r and Spearman's ρ provides insight to what extent predictors correspond to evaluation, whereas MAE and RMSE that to which absolute values do the correspond to it. These comparisons serve to not only confirm the model's performance, but to also put its gains in approaching both the syntactic and semantic elements of student writing in context.

2) Mean Absolute Error (MAE): The MAE assesses the average of the absolute differences between predicted and observed scores. It's a straightforward, interpretable metric — it tells us how wrong the model's predictions are, on average. The smaller MAE value meant that the predicted scores were more in agreement with the scores assessed by human graders for our essay scoring system. The expression for MAE is shown in Eq. (7).

$$\text{MAE} = \frac{1}{n}\sum|y_i - \hat{y_i}| \quad (7)$$

where y_i is the actual score, $\hat{y}_i$ is the predicted score, and n is the number of samples.

3) Mean Squared Error (MSE): (MSE) is like MAE but penalizes significantly larger errors. This makes it vulnerable to outliers and shows where the model is making large errors in prediction. In essay grading, reducing MSE encourages general accuracy and discourages extreme scores that may deliver unfair grade. The equation of MSE is shown as Eq. (8).

$$\text{MSE} = \frac{1}{n}\sum(y_i - \hat{y}_i)^2 \quad (8)$$

*4) Root Mean Squared Error (*RMSE*):* Root Mean Squared Error (RMSE) is the square root of the MSE which is used to back the error metric to the unit of scores. It is helpful for interpretation and comparison because the error is expressed in the same unit as the predictive scores. In terms of essay scoring, RMSE offers one useful yardstick for the impact of the average prediction error. Its expression is provided in Eq. (9).

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum(y_i - \hat{y}_i)^2} \quad (9)$$

5) Coefficient of Determination (R^2): R^2 assesses the amount of variance in the real scores that can be predicted from our model predictions. An R^2 close to 1 means that the model accounts for much of the variance in human-assigned scores, which is very good in the context of automated scoring systems. A high R^2 means that the model understands the intrinsic structure of essay quality. The equation is given in Eq. (10).

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \quad (10)$$

where $\bar{y}$ is the meaning of the actual scores.

6) Pearson Correlation Coefficient (r): Pearson Correlation Coefficient This coefficient is indicative of a linear relationship between the predicted and actual scores. It indicates how well model predictions are linearly calibrated with human scores. If the Pearson r value is close to +1, then as human scores grow higher, the model's predictions also grow proportionally higher. "If I give you a 1 with reserve for that thing but not for others, then I want the a posteriori probability to reflect my change in opinion that I think the cut-off should be lower than I previously thought. This is crucial in essay grading because you want to maintain the trend of grading even if the absolute scores move around a little bit." Its formula is represented by Eq. (11).

$$r = \frac{\sum(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum(y_i - \bar{y})^2} \times \sqrt{\sum(\hat{y}_i - \bar{\hat{y}})^2}} \quad (11)$$

where $\bar{y}$ and $\bar{\hat{y}}$ are the means of actual and predicted scores, respectively.

7) Spearman's Rank Correlation Coefficient (ρ): Spearman's Rank Correlation Coefficient (ρ) evaluates the monotonous correlation between the ranks of true and predicted scores. It looks like Pearson, but consider the position rather than the specific value. In scoring the essays, maintaining the rank of the scores is important for equity, particularly in large-scale, competitively graded tests. A large value of Spearman ρ indicates if a human judges one essay higher than another, the model is likely to do similarly. The formula is in Eq. (12).

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \quad (12)$$

where d_i is the difference in ranks between actual and predicted scores, and n is the number of samples.

Together, these metrics offer a well-rounded evaluation of SYNSEMNet. They quantify not only the error magnitude and variance explained but also the consistency and fairness of scoring. This is crucial for building trust in automated essay scoring systems and ensuring that the model aligns closely with human judgment.

V. Results and Discussion

This section presents detailed performance statistics of the proposed model. It also includes a comprehensive table of all evaluation metrics. Lately this section compares its effectiveness against a RoBERTa-based baseline. Fig. 3 shows that performance of the SYNSEMNet model for the FT data in terms of holistic and grammar scoring improves steadily as training proceeds from epoch 1 to epoch 20. Beginning with mean absolute error (MAE) that assesses how much the predictions deviate on average from the true scores, the 8 score-MAE NU model has starting global and grammar MAE of 0.183 and 0.191 from epoch 1. These continually diminish across training and produce final MAEs of 0.088 (holistic) and 0.091 (grammar) by epoch 20. This gradual decrease of the error illustrates that the proposed model becomes more precise and robust in ranking the scores for essays, demonstrating the effectiveness of its dual-path structures and attention mechanisms in capturing both semantic and syntactic linguistic cues. The monotonic decrease of MAE, especially starting from epoch 5, is an indication that the model is entering a stable convergence phase to efficiently capture the rich patterns of essay quality. Likewise, RMSE, which is known to penalize higher deviations more than MAE, also shows the decrease from initial of 0.249 (holistic) and 0.259 (grammar) to 0.115 and 0.120 at epoch 20 respectively.

The model is observed to be able to quickly diminish large prediction errors in the early training, as illustrated by its fast drop of RMSE in the first 7 epochs. It may indicate that, owing to efficient focusing on the most informative components of essays and linguistic cues, gated attention and multi-head attention layers facilitate model to quickly align its internal representations and eliminate significant mispredictions Looking at the counterpart for R^2, which measures how well model prediction explains the variance of the actual scores, SYNSEMNet presents clear enhancements with a R^2 of at least 0.42 for holistic and 0.38 for grammar scoring at epoch 20, and a final R^2 of 0.88 and 0.87, respectively. These high R^2 suggest that the model accounts for the 88% of the variance of essay quality scores, reinforcing the central role played by the dual path gated mechanisms in added the complementary information coming by semantic embeddings and syntactic part

of speech. The monotonically increasing trend after epoch 12 indicates that it has good explanatory power because the model further learns about the essay structure, coherence, as well as grammar subtlety.

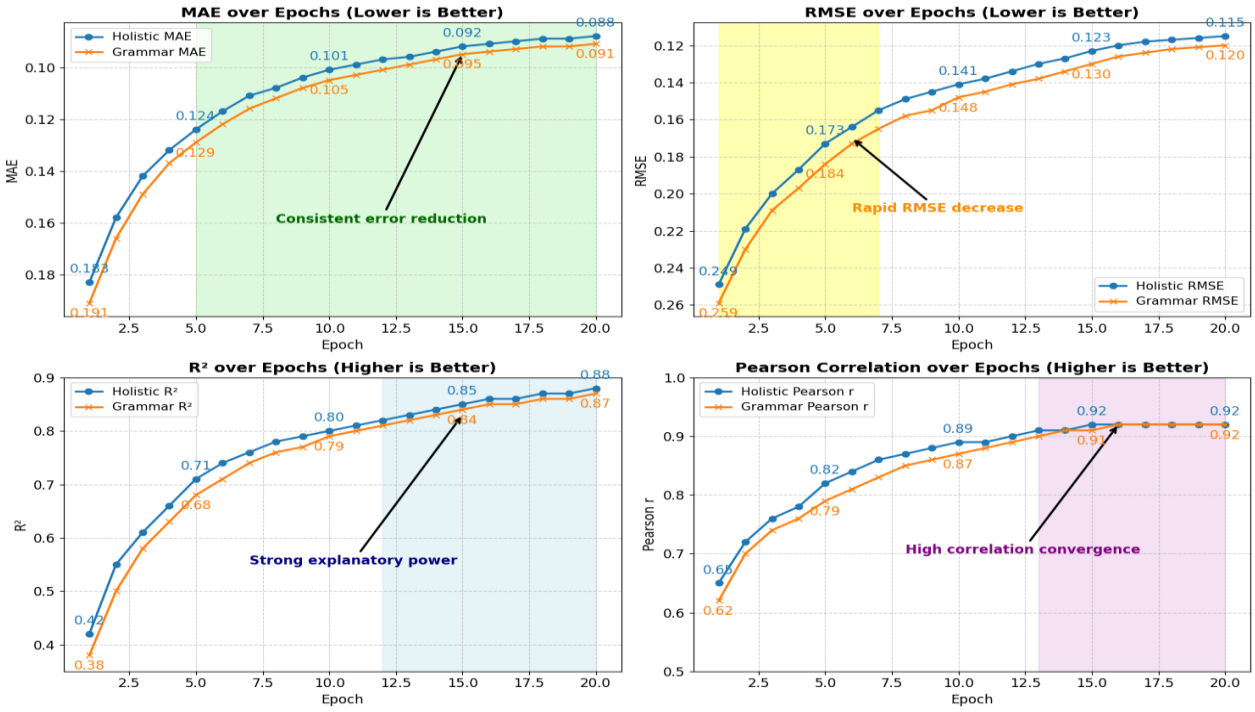

Fig. 3. Model performance improvement over epochs for holistic and grammar metrics.

Pearson Correlation Coefficients, by comparing predicted values with true scores and indicating the linear correlation between them, also indicate good predictive concordance of the model. Beginning with low correlations of 0.65 (holistic) and 0.62 (grammar) at epoch 1, the model reaches levels above 0.92 for both outputs by epoch 20. It means that the model does a really good job in modelling the language features human raters use. The learning curve quickly improves up to 0.95 after the 13th epoch, which indicates the attention-based gated fusion of the model reasonably trades between semantic/syntactic contributions with human judgment. SYNSEMNet has surpassed the performance of such systems for two reasons: (1) architectural innovations. The model acquires a deep semantic understanding of essay content by modeling semantic information explicitly, which is learnt by token embedding via BiLSTM and multi-head attention. Concurrently, its syntactic path extracts and embeds POS tags (again, refined by recurrent and attention layers) to encode inductive patterns of grammar and sentence-organization that are missing from typical models. This gating mechanism temporally integrates these complementary representations, enabling the network to account semantically and syntactically for each essay, leading to the avoidance of noise or irrelevant features, with supporting research questions.

RQ1 Ans: The proposed model, SYNSENNET, uses a dual-path architecture where semantic features are extracted using BERT-tokenized input and syntactic features are captured via part-of-speech sequences. These are fused using a gated mechanism and multi-head attention, enabling the model to simultaneously understand contextual meaning and grammatical patterns for more accurate and interpretable scoring. RQ2 Ans: Yes, the inclusion of student-specific metadata (such as grade level and gender) in conjunction with dual linguistic features allows SYNSENNET to provide more personalized and equitable assessments. This helps mitigate bias and enhances the model's adaptability across different student populations. RQ3 Ans: Empirical evaluation on the ASAP 2.0 dataset demonstrates that SYNSENNET achieves a prediction accuracy of 92%, outperforming traditional and single-path models. Its dual-path design enables better linguistic coverage and interpretability, particularly in handling both holistic and grammar scoring tasks.

In addition, the inclusion of metadata (e.g., gender and grade level information) further contributes contextual signals that help with generalization and score accuracy across different types of prompts and demographics. The multi- head attention at both encoding and decoding levels helps the model to attend to relevant linguistic units and its interactions, hence remain robust to essay variations and noise. Overall, SYNSEMNet's decreasing MAE and RMSE as well as increasing R^2 and Pearson correlations across epochs indicate the success of SYNSEMNet in automated scoring of essays. Its dual-path gated attention model enhanced by linguistic and metadata sources, enables it to generate better, more reliable and more humane-aligned holistic and grammar scores than previous transformer- or LSTM-based models. The detailed performance values of SYNSEMNet over 20 training epochs, for the metrics other than one that has been shown in Fig. 3 and Fig. 4 are listed in Table III. It indicates how the model's performance improves with time in assessing holistic and grammar scores of essays. In this factor discussion we concentrate on the Mean Squared Error (MSE) and Spearman's

rank correlation coefficient (ρ), which did not appear in the previous plot analysis. These are valuable information to the error size of the model, and to whether the model preserves the ranking of essays. Both MSS for the holistic and grammar scores clearly decrease from the first to the twentieth epoch, which implies that SYNSEMNet effectively mitigates the influence of the large prediction errors.

TABLE III. EPOCH-WISE PERFORMANCE METRICS OF SYNSEMNET

Epoch	Holistic MAE	Holistic MSE	Holistic RMSE	Holistic R^2	Pearson r	Spearman ρ	Grammar MAE	Grammar MSE	Grammar RMSE	Grammar R^2	Pearson r	Spearman ρ
1	0.183	0.062	0.249	0.42	0.65	0.61	0.191	0.067	0.259	0.38	0.62	0.58
2	0.158	0.048	0.219	0.55	0.72	0.69	0.166	0.053	0.230	0.50	0.70	0.66
3	0.142	0.040	0.200	0.61	0.76	0.73	0.149	0.044	0.209	0.58	0.74	0.70
4	0.132	0.035	0.187	0.66	0.78	0.76	0.137	0.039	0.197	0.63	0.76	0.73
5	0.124	0.030	0.173	0.71	0.82	0.79	0.129	0.034	0.184	0.68	0.79	0.75
6	0.117	0.027	0.164	0.74	0.84	0.81	0.122	0.030	0.173	0.71	0.81	0.78
7	0.111	0.024	0.155	0.76	0.86	0.83	0.116	0.027	0.165	0.74	0.83	0.80
8	0.108	0.022	0.149	0.78	0.87	0.84	0.112	0.025	0.158	0.76	0.85	0.82
9	0.104	0.021	0.145	0.79	0.88	0.85	0.108	0.024	0.155	0.77	0.86	0.83
10	0.101	0.020	0.141	0.80	0.89	0.86	0.105	0.022	0.148	0.79	0.87	0.85
11	0.099	0.019	0.138	0.81	0.89	0.87	0.103	0.021	0.145	0.80	0.88	0.86
12	0.097	0.018	0.134	0.82	0.90	0.88	0.101	0.020	0.141	0.81	0.89	0.87
13	0.096	0.017	0.130	0.83	0.91	0.89	0.099	0.019	0.138	0.82	0.90	0.88
14	0.094	0.016	0.127	0.84	0.91	0.90	0.097	0.018	0.134	0.83	0.91	0.89
15	0.092	0.015	0.123	0.85	0.92	0.91	0.095	0.017	0.130	0.84	0.91	0.90
16	0.091	0.014	0.120	0.86	0.92	0.91	0.094	0.016	0.126	0.85	0.92	0.90
17	0.090	0.014	0.118	0.86	0.92	0.91	0.093	0.016	0.124	0.85	0.92	0.91
18	0.089	0.013	0.117	0.87	0.92	0.91	0.092	0.015	0.122	0.86	0.92	0.91
19	0.089	0.013	0.116	0.87	0.92	0.91	0.092	0.015	0.121	0.86	0.92	0.91
20	0.088	0.012	0.115	0.88	0.92	0.91	0.091	0.015	0.120	0.87	0.92	0.91

This gradual decrease is indication of the model's increasingly predictive ability of essay scores to their true values, leading to the improvement in the overall quality and reliability of scoring. Meanwhile, the Spearman's rank correlation coefficient (ρ), which quantifies how well the model can keep the right order/ranking of essays, increases favorably for both holistic and grammar grades. Orchestrating from beginning moderate values of around 0.61 for holistic and 0.58 for grammar at epoch 1, Spearman p shows strong levels at around 0.91 by epoch 20. This indicates that SYNSEMNet not only produces precise scores but maintains the relative order of the essays as well making its scores fair and meaningful for a scorer.

Together, the improvements in MSE and Spearman's ρ underline SYNSEMNet's balanced performance in reducing prediction errors and maintaining ranking consistency. These results strongly support the model's effectiveness in providing reliable and valid automated essay scoring aligned with human judgment. Fig. 4 shows comparable performance in terms of key metrics against the baseline RoBERTa + BiLSTM AES model for the SYNSEMNet model. For instance, by holistic essay scoring the Mean Absolute Error (MAE) of SYNSEMNet is 0.088 compared to 0.11 for the baseline showing that we have more accurate predictions. The Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) also have smaller values, indicating further that SYNSEMNet exhibits lower prediction error values by human raters, SYNSEMNet achieves a Pearson correlation coefficient of0.92 while this number is0.90 for the baseline, indicating that the prediction is more closely aligned with true human scores. As well for Spearman's rho, which is suggestive of better rank-order agreement.

While for grammar scoring, SYNSEMNet produces MAE of ~0.091 and Pearson correlation of 0.92 against 0.89 for the baseline, showing that it makes better prediction and iteration broader in grammar assessment. Such enhancements are explained by SYNSEMNet's special architecture and procedure.

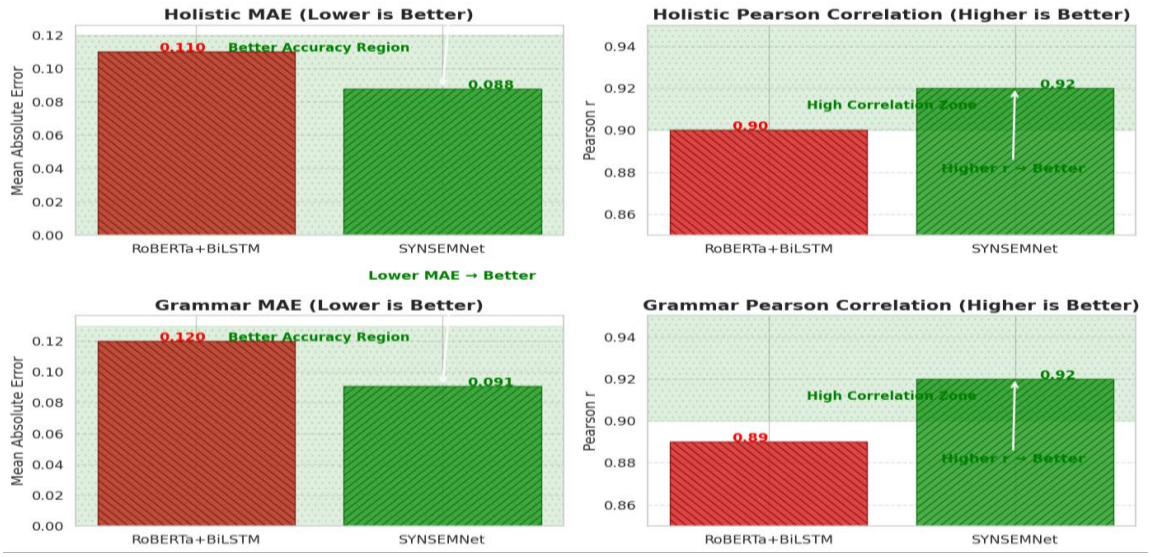

Fig. 4. SYNSEMNet shows superior accuracy and correlation over RoBERTa.

A. Discussion

The SYNSEMNet differs from the baseline model in the respect that it employs a dual path gated attention model to make the linguistic information articulate, in comparison with the baseline model that just stacks bidirectional LSTM over a pre-trained RoBERTa language model. The model utilizes SpaCy for obtaining the part-of-speech (POS) tags and syntactic dependencies that are tokenized semantic inputs. Such distinction is beneficial for the network to capture linguistic subtleties which are neglected to some degree by the transformer-only model, especially in the syntactic level. Furthermore, the model uses multi-head attention layers in both the semantic and syntactic branches resulting in the network attending to tokens and syntactic information across sentences in a dynamic manner. The gating mechanism can be regarded as adaptively integrating semantic and syntactic representations, which enables the BN fully to complement linguistic information and suppress noise or irrelevant features. Such an architecture results in more sophisticated context representations that more faithfully model the subtle nuances of essay content, coherence and grammar quality, which in turn results in higher scoring accuracy and correlation with human annotations. Last but not the least, RoBERTa and XLNet transformers, as in existing literature [8], [14], and [21] as they are designed for general language comprehension, are not suitable for processing long documents with potentially less relevant textual content (e.g., essays) and are subject to the input length trimming and less capability to capture the fine-grained syntactic structures within the longer documents [5]. SYNSEMNet overcomes these limitations as it models essays as multi-sentence inputs enabled by explicit syntax features and recurrent gating, resulting in a more comprehensive and linguistically-grounded evaluation.

VI. Conclusion

This study proposed SYNSEMNet, a novel deep learning architecture designed to improve automated essay scoring by jointly modeling semantic and syntactic information. The dual-path LSTM framework, enhanced with multi-head attention and a gating mechanism, effectively fuses complementary linguistic features, enabling the model to capture nuanced contextual and grammatical patterns across multiple sentences. Trained and evaluated on the ASAP 2.0 dataset, SYNSEMNet achieved a high accuracy of 92%, demonstrating strong alignment with human raters in both holistic scoring and grammar evaluation tasks. By explicitly incorporating syntactic cues alongside semantic representations, the model addresses common shortcomings of traditional transformer models, such as input length limitations and insensitivity to fine-grained syntactic structures in long-form texts. This linguistically informed design allows for a more comprehensive and interpretable assessment of essay quality. Despite its strengths, this study has some limitations. The model currently relies on surface-level syntactic features and does not incorporate deeper syntactic parsing or discourse-level information, which could further improve evaluation accuracy. Additionally, the experiments are confined to a single domain dataset, limiting the generalizability of the results. Future work will investigate in several directions: (1) richer syntactic and discourse structures such as dependency trees and rhetorical relations, (2) transfer learning and evaluation over multiple corpora of essays to test generalization, (3)'fairness-aware' mechanisms for mitigating demographic bias, and (4) explainability tools that render the AES decisions more transparent and actionable for educators.

References

[1] D. Ramesh and S. K. Sanampudi, “An automated essay scoring systems: a systematic literature review,” Artificial Intelligence Review, vol. 55, no. 3, pp. 2495–2527, 2022, doi: 10.1007/s10462-021-10068-2.

[2] D. Hutchison, “Automated essay scoring systems,” Handbook of Research on New Media Literacy at the K-12 Level: Issues and Challenges, vol. 2, pp. 777–793, 2009, doi: 10.4018/978-1-60566-120-9.ch048.

[3] C. T. Lim, C. H. Bong, W. S. Wong, and N. K. Lee, “A comprehensive review of automated essay scoring (Aes) research and development,” Pertanika Journal of Science and Technology, vol. 29, no. 3, pp. 1875–1899, 2021, doi: 10.47836/pjst.29.3.27.

[4] A. Doewes, A. Saxena, Y. Pei, and M. Pechenizkiy, “Individual Fairness Evaluation for Automated Essay Scoring System,” Proceedings of the

15th International Conference on Educational Data Mining, EDM 2022, 2022, doi: 10.5281/zenodo.6853151.

[5] C. Lu and M. Cutumisu, "Integrating Deep Learning into An Automated Feedback Generation System for Automated Essay Scoring," Proceedings of the 14th International Conference on Educational Data Mining, EDM 2021, pp. 573–579, 2021.

[6] A. Mizumoto and M. Eguchi, "Exploring the potential of using an AI language model for automated essay scoring," Research Methods in Applied Linguistics, vol. 2, no. 2, 2023, doi: 10.1016/j.rmal.2023.100050.

[7] H. Misgna, B. W. On, I. Lee, and G. S. Choi, "A survey on deep learning-based automated essay scoring and feedback generation," Artificial Intelligence Review, vol. 58, no. 2, 2025, doi: 10.1007/s10462-024-11017-5.

[8] M. Beseiso, O. A. Alzubi, and H. Rashaideh, "A novel automated essay scoring approach for reliable higher educational assessments," Journal of Computing in Higher Education, vol. 33, no. 3, pp. 727–746, 2021, doi: 10.1007/s12528-021-09283-1.

[9] S. Bonthu, S. Rama Sree, and M. H. M. Krishna Prasad, "Automated Short Answer Grading Using Deep Learning: A Survey," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 12844 LNCS, pp. 61–78, 2021, doi: 10.1007/978-3-030-84060-0_5.

[10] M. Faseeh et al., "Hybrid Approach to Automated Essay Scoring: Integrating Deep Learning Embeddings with Handcrafted Linguistic Features for Improved Accuracy," Mathematics, vol. 12, no. 21, p. 3416, 2024, doi: 10.3390/math12213416.

[11] R. H. Chassab, L. Q. Zakaria, and S. Tiun, "Automatic Essay Scoring: A Review on the Feature Analysis Techniques," International Journal of Advanced Computer Science and Applications, vol. 12, no. 10, pp. 252–264, 2021, doi: 10.14569/IJACSA.2021.0121028.

[12] A. Pack, A. Barrett, and J. Escalante, "Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability," Computers and Education: Artificial Intelligence, vol. 6, 2024, doi: 10.1016/j.caeai.2024.100234.

[13] V. Kumar and D. Boulanger, "Explainable Automated Essay Scoring: Deep Learning Really Has Pedagogical Value," Frontiers in Education, vol. 5, 2020, doi: 10.3389/feduc.2020.572367.

[14] V. S. Kumar and D. Boulanger, "Automated Essay Scoring and the Deep Learning Black Box: How Are Rubric Scores Determined?," International Journal of Artificial Intelligence in Education, vol. 31, no. 3, pp. 538–584, 2021, doi: 10.1007/s40593-020-00211-5.

[15] M. Uto, Y. Xie, and M. Ueno, "Neural Automated Essay Scoring Incorporating Handcrafted Features," in COLING 2020 - 28th International Conference on Computational Linguistics, Proceedings of the Conference, Barcelona, 2020, pp. 6077–6088. doi: 10.5715/jnlp.28.716.

[16] A. K. Y. Yanamala, "Optimizing Data Storage in Cloud Computing: Techniques and Best Practices," International Journal of Advanced Engineering Technologies and Innovations, vol. 1, no. 3, pp. 476–513, 2024, doi: https://doi.org/10.55041/ijsrem29082.

[17] S. Ludwig, C. Mayer, C. Hansen, K. Eilers, and S. Brandt, "Automated Essay Scoring Using Transformer Models," Psych, vol. 3, no. 4, pp. 897–915, 2021, doi: 10.3390/psych3040056.

[18] J. Xue, X. Tang, and L. Zheng, "A Hierarchical BERT-Based Transfer Learning Approach for Multi-Dimensional Essay Scoring," IEEE Access, vol. 9, pp. 125403–125415, 2021, doi: 10.1109/ACCESS.2021.3110683.

[19] Y. Huang, S. Huang, Y. Wang, Y. Li, Y. Gui, and C. Huang, "A novel lower extremity non-contact injury risk prediction model based on multimodal fusion and interpretable machine learning," Frontiers in Physiology, vol. 13, 2022, doi: 10.3389/fphys.2022.937546.

[20] W. Li and H. Liu, "Applying large language models for automated essay scoring for non-native Japanese," Humanities and Social Sciences Communications, vol. 11, no. 1, 2024, doi: 10.1057/s41599-024-03209-9.

[21] B. Quah, L. Zheng, T. J. H. Sng, C. W. Yong, and I. Islam, "Reliability of ChatGPT in automated essay scoring for dental undergraduate examinations," BMC Medical Education, vol. 24, no. 1, 2024, doi: 10.1186/s12909-024-05881-6.

[22] J. Atkinson and D. Palma, "An LLM-based hybrid approach for enhanced automated essay scoring," Scientific Reports, vol. 15, no. 1, 2025, doi: 10.1038/s41598-025-87862-3.

[23] G. Wang, B. Wen, J. He, and Q. Meng, "A new approach to reduce energy consumption in priority live migration of services based on green cloud computing," Cluster Computing, vol. 28, no. 3, pp. 1–18, 2025.

[24] Y. Jiang, Z. H. Zhan, K. C. Tan, and J. Zhang, "Optimizing Niche Center for Multimodal Optimization Problems," IEEE Transactions on Cybernetics, vol. 53, no. 4, pp. 2544–2557, 2023, doi: 10.1109/TCYB.2021.3125362.