Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.081004

ARTICLE

A Multi-Branch Transformer-Enhanced Neural Framework for Joint Morphological Representation Learning

Laura Baitenova[1], Gulnar Mukhamejanova[2], Gauhar Munaitbas[3,*], Saken Mambetov[1] and Zhanna Mukanova[1]

[1]Higher School of Information Technology, Turan University, Almaty, Kazakhstan
[2]School of Digital Technologies, Narxoz University, Almaty, Kazakhstan
[3]Home Credit Bank JSC, Almaty, Kazakhstan

*Corresponding Author: Gauhar Munaitbas. Email: gmunaitbas@gmail.com

Received: 21 February 2026; Accepted: 21 May 2026; Published: 23 July 2026

ABSTRACT: Morphological parsing is a fundamental task in natural language processing, particularly for morphologically rich languages where words encode complex grammatical and semantic information. This paper proposes a multi-branch Transformer-enhanced neural framework for joint morphological representation learning, designed to improve segmentation and classification accuracy by integrating complementary feature extraction mechanisms. The proposed architecture combines convolutional layers for capturing local morphological patterns, recurrent layers for modeling sequential dependencies, and Transformer-based self-attention for learning global contextual relationships. This hybrid design enables the model to generate robust and context-aware representations that enhance morphological understanding. The framework is trained using morphologically annotated datasets and evaluated using standard performance metrics, including F1-score and classification accuracy. Experimental results demonstrate that the proposed model significantly outperforms conventional and single-architecture baselines in both segmentation and morphological classification tasks. The learned representations exhibit strong discriminative capability, allowing accurate identification of morpheme boundaries and grammatical features. Furthermore, the model demonstrates stable convergence behavior and strong generalization performance across diverse linguistic conditions. These findings confirm the effectiveness of integrating multi-level contextual and structural feature extraction mechanisms, establishing the proposed framework as a robust and scalable solution for advanced morphological parsing and representation learning in modern natural language processing applications.

KEYWORDS: Morphological parsing; transformer architecture; multi-branch neural networks; morphological segmentation; contextual representation learning; natural language processing; morphological classification; deep learning; self-attention mechanism

1 Introduction

Morphological analysis constitutes a fundamental component of natural language processing, as it enables the decomposition of words into their constituent linguistic units and the extraction of grammatical, syntactic, and semantic features necessary for higher-level language understanding tasks. Accurate morphological parsing facilitates downstream applications such as machine translation, information extraction, syntactic parsing, and semantic analysis, particularly in morphologically rich languages where a single word may encode complex grammatical information. Traditional morphological parsing methods relied heavily on rule-based systems and manually engineered features, which, although effective in constrained

environments, suffer from limited scalability, reduced adaptability, and poor generalization across domains and languages [1]. These limitations have motivated the adoption of machine learning and, more recently, deep learning techniques that can automatically learn representations directly from data.

The emergence of neural network architectures has fundamentally transformed morphological analysis by enabling the modeling of sequential dependencies and contextual relationships within text. Recurrent neural networks, particularly Long Short-Term Memory (LSTM) models, have demonstrated strong capabilities in capturing temporal and sequential dependencies due to their ability to maintain long-range contextual information through gated memory mechanisms [2]. Such architectures have proven effective for tasks including part-of-speech tagging, lemma prediction, and morphological feature classification. However, despite their strengths, recurrent models process tokens sequentially, which inherently limits their ability to fully capture global contextual relationships and increases computational latency, especially for long sequences.

To address these limitations, attention-based architectures have gained widespread adoption. The Transformer model, which relies entirely on self-attention mechanisms, enables the modeling of both local and global dependencies simultaneously by allowing each token to attend to all other tokens in the sequence [3]. This capability significantly enhances contextual representation learning and improves performance across a wide range of NLP tasks. Transformer-based models have demonstrated superior effectiveness in capturing complex syntactic and semantic relationships, making them particularly suitable for morphological parsing, where contextual cues often determine grammatical interpretation [4]. Nevertheless, purely Transformer-based architectures may overlook certain fine-grained sequential patterns that recurrent networks capture naturally, particularly in low-resource or highly inflectional language settings.

Recent research has highlighted the importance of combining complementary computational approaches to leverage their respective strengths. Hybrid neural frameworks that integrate rule-based methods with neural architectures, including transformer-based models and probabilistic sequence labeling techniques, can effectively capture both morphological structure and contextual dependencies, thereby improving robustness in low-resource NLP tasks [5]. Such hybridization enables the extraction of hierarchical linguistic features, improves generalization across diverse linguistic structures, and enhances prediction accuracy. Furthermore, multi-branch architectures have emerged as a powerful design paradigm, allowing different network components to specialize in learning distinct aspects of linguistic representation while contributing to a unified predictive framework [6]. This modular approach enhances feature diversity and reduces the risk of information loss during representation learning.

Another critical advancement in morphological analysis is the adoption of multi-task learning strategies, which enable the simultaneous prediction of multiple linguistic features such as lemmas, part-of-speech tags, and grammatical attributes. Multi-task learning facilitates shared representation learning, allowing the model to capture common underlying structures across related tasks while improving generalization and reducing overfitting [7]. This approach is particularly beneficial in morphological parsing, where different linguistic features are inherently interconnected and mutually informative. By jointly optimizing multiple objectives, multi-task architectures enhance the overall consistency and accuracy of morphological predictions.

Despite these advances, challenges remain in effectively integrating contextual and sequential representations while maintaining architectural efficiency and prediction accuracy. Existing models often struggle to balance local sequential modeling with global contextual understanding, resulting in suboptimal performance in complex linguistic environments. Furthermore, insufficient feature fusion mechanisms may limit the ability of models to fully exploit complementary information derived from different architectural components [8]. These challenges highlight the need for novel neural architectures capable of effectively

combining contextual attention mechanisms with sequential representation learning in a unified and scalable framework.

To address these limitations, this paper proposes a multi-branch Transformer-enhanced neural framework for joint morphological representation learning. The proposed architecture integrates a Transformer-based contextual encoder with a recurrent sequential encoder and a feature fusion module to capture both global and local linguistic dependencies. The model employs a multi-task decoding strategy to simultaneously predict lemmas, part-of-speech tags, and grammatical features, enabling comprehensive morphological analysis. This integrated approach enhances representation richness, improves predictive accuracy, and provides a scalable solution for contextual morphological parsing in modern NLP applications. In addition, a multi-task learning strategy is adopted to jointly predict morphological segmentation, lemma forms, part-of-speech tags, and grammatical attributes within a single unified framework.

The main contributions of this work are summarized as follows:

- A multi-branch neural architecture that integrates CNN, LSTM, and Transformer components to jointly capture local morphological patterns, sequential dependencies, and global contextual information.
- A unified multi-task learning framework for simultaneous morpheme segmentation and morphological classification, enabling shared feature representation and improved generalization.
- A comprehensive experimental evaluation across multilingual datasets, demonstrating consistent improvements over traditional and deep learning-based baselines.
- An in-depth analysis, including ablation studies and interpretability evaluation, to assess the contribution of each architectural component and provide insights into model behavior.

Unlike prior studies that treat segmentation and classification as separate tasks or rely on single-model architectures, the proposed approach focuses on the integration of complementary feature extraction mechanisms within a unified framework, improving both performance and model consistency.

The remainder of this paper is organized as follows. Section 2 reviews related works in morphological parsing and neural representation learning. Section 3 describes the proposed methodology, including the overall architecture and Transformer block design. Section 4 presents the experimental setup and results, along with quantitative and qualitative analyses. Section 5 discusses the findings and implications of the study. Finally, Section 6 concludes the paper and outlines directions for future research.

2 Related Works

Morphological parsing has traditionally been approached through rule-based and statistical frameworks that rely on handcrafted linguistic rules, morphological lexicons, and probabilistic models. These early systems were designed to explicitly encode grammatical knowledge and achieved satisfactory performance in narrowly defined linguistic settings. However, their dependence on expert-defined rules limited their scalability and made adaptation to new languages and domains costly and time-consuming. As linguistic variability increased and multilingual NLP applications became more prevalent, these approaches struggled to generalize beyond curated datasets and controlled environments [9]. The growing availability of large annotated corpora motivated a shift toward data-driven learning paradigms capable of automatically extracting morphological patterns from raw text. This transition laid the groundwork for neural approaches, which significantly reduced the need for manual feature engineering while improving robustness to linguistic variation [10].

The introduction of neural network models marked a turning point in morphological analysis by enabling end-to-end learning of linguistic representations. Early neural approaches primarily employed feedforward architectures that learned distributed word representations and morphological features directly

from data. While these models improved generalization and reduced reliance on handcrafted features, they lacked mechanisms to model sequential dependencies across tokens, which are essential for resolving contextual ambiguity in morphological parsing [11]. Recurrent neural networks addressed this limitation by introducing memory-based architectures capable of processing sequences token by token. Long Short-Term Memory networks, in particular, demonstrated strong performance by preserving relevant contextual information through gated memory units, enabling more accurate morphological tagging and lemma prediction [12].

Bidirectional recurrent architectures further enhanced morphological parsing by capturing both preceding and succeeding contextual information within a sentence. By processing sequences in forward and backward directions, bidirectional LSTM models improved the representation of contextual dependencies and significantly enhanced prediction accuracy in sequence labeling tasks [13]. These models became a dominant paradigm for morphological analysis, particularly in morphologically rich languages. Nevertheless, their inherently sequential processing nature limited computational efficiency and hindered scalability. Additionally, recurrent models often struggled to capture global contextual relationships when dependencies spanned long distances, motivating the exploration of alternative architectures capable of modeling long-range interactions more effectively [14].

Attention mechanisms emerged as a powerful solution to these limitations by allowing models to dynamically focus on relevant parts of the input sequence. Attention-based models improved contextual representation learning by enabling direct interactions between distant tokens, thereby enhancing the modeling of syntactic and semantic relationships [15]. This concept was fully realized in the Transformer architecture, which relies exclusively on self-attention mechanisms and eliminates recurrence entirely. Transformer-based models enable efficient parallel computation and facilitate the capture of global contextual dependencies across entire sequences, leading to substantial performance gains in a wide range of NLP tasks, including morphological parsing [16]. Their ability to generate rich contextual embeddings has proven particularly beneficial for resolving morphological ambiguity [17].

Despite their strengths, purely Transformer-based architectures may overlook certain fine-grained sequential patterns that recurrent models naturally capture. This observation has led to increasing interest in hybrid architectures that combine recurrent and attention-based components. Such hybrid models exploit the complementary strengths of sequential memory and contextual attention, producing more expressive and robust linguistic representations [18]. Transformer-LSTM hybrid frameworks have demonstrated improved performance in sequence labeling and morphological analysis by jointly modeling local token order and global sentence context [19]. These findings suggest that integrating multiple architectural paradigms can yield richer representations than relying on a single modeling approach [20].

More recently, multi-branch and multi-task neural architectures have gained prominence in morphological parsing research. Multi-branch designs allow different network components to specialize in learning complementary features, which are subsequently integrated through feature fusion mechanisms [21]. In parallel, multi-task learning frameworks enable the joint prediction of lemmas, part-of-speech tags, and grammatical features within a unified model, promoting shared representation learning and improving generalization [22]. Contextualized word representations further enhance these frameworks by dynamically adapting embeddings to linguistic context, thereby improving robustness and accuracy [23]. Advances in regularization and optimization techniques have also contributed to the development of deeper and more stable models [24]. Nevertheless, challenges remain in effectively fusing contextual and sequential representations and ensuring efficient information sharing across tasks. These limitations motivate the development of advanced multi-branch Transformer-enhanced architectures for joint morphological representation learning.

3 Materials and Methods

This study proposes a multi-branch Transformer-enhanced neural framework for joint morphological representation learning, designed to address the challenges associated with accurate segmentation and classification of morphological structures in morphologically rich languages. The methodological pipeline begins with the acquisition of raw textual data, followed by systematic preprocessing procedures that ensure consistency, structural integrity, and compatibility with neural processing requirements. Preprocessing includes normalization of textual symbols, tokenization into word-level units, and optional subword segmentation to improve representation of rare or morphologically complex forms. The processed data are then integrated with annotated morphological datasets containing ground-truth information such as morpheme boundaries, part-of-speech tags, and grammatical feature labels. These data serve as the foundation for training and validating the proposed model. The neural framework leverages Transformer-based contextual encoding combined with joint representation learning to capture both local morphological patterns and global contextual dependencies. The model is trained using supervised learning with multi-task objectives, allowing simultaneous optimization of segmentation and classification tasks. Model performance is quantitatively assessed using established evaluation metrics, including F1-score, classification accuracy, and loss convergence behavior, ensuring comprehensive evaluation of both structural and semantic prediction capabilities.

Fig. 1 illustrates the initial stages of the methodological pipeline, focusing on text preprocessing and dataset preparation, which are essential for enabling effective morphological representation learning. The process begins with the input text, denoted as a sequence of characters forming a sentence

$$\mathrm{S} = \{\mathrm{c}_1, \mathrm{c}_2, \ldots, \mathrm{c}_\mathrm{L}\}, \tag{1}$$

where L represents the total number of characters. The first preprocessing step involves normalization, which transforms the raw character sequence into a standardized form by removing inconsistencies such as case variations, irregular spacing, and encoding artifacts. Formally, the normalization function can be defined as a transformation

$$\mathrm{N} : \mathrm{S} \rightarrow \mathrm{S}', \tag{2}$$

where S' is the normalized character sequence. This step ensures that semantically equivalent tokens are represented consistently, thereby reducing noise and improving model generalization.

Following normalization, the normalized sequence S' is segmented into a sequence of tokens using a tokenization function, resulting in a token sequence

$$W = \{w_1, w_2, \ldots, w_n\} \tag{3}$$

where n denotes the number of tokens in the sentence. Each token w_i represents a linguistic unit such as a word or subword. To further enhance the model's ability to handle morphologically complex and rare words, optional subword segmentation may be applied. This operation decomposes each token w_2 into a sequence of subword units

$$w_i = \{s_{i,1}, s_{i,2}, \ldots, s_{i,k}\}, \tag{4}$$

where k_i denotes the number of subword components. Subword decomposition improves representation learning by allowing the model to capture internal morphological structure, particularly prefixes, roots, and suffixes.

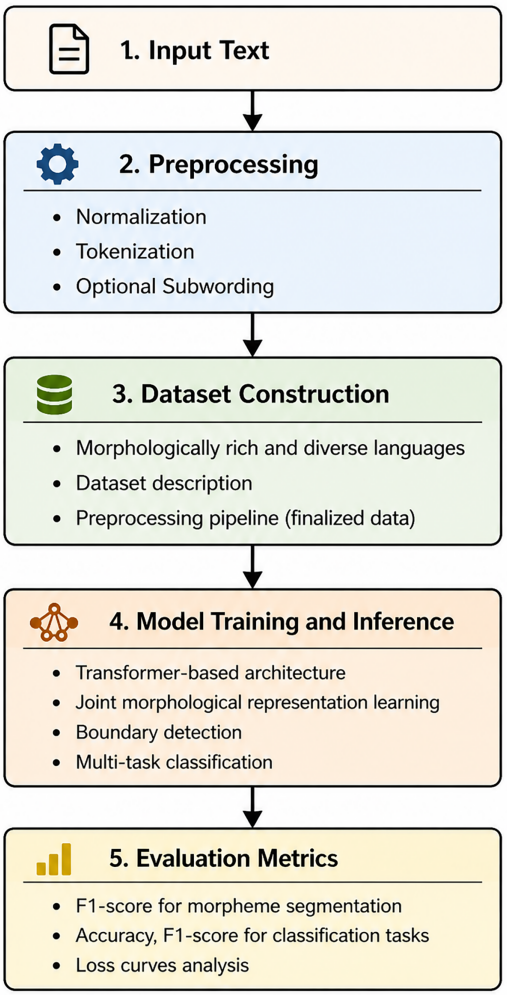

Figure 1: Methodological pipeline of the proposed framework for joint morphological representation learning and evaluation.

Once tokenization and optional subword segmentation are complete, each token or subword unit is mapped to a numerical representation suitable for neural processing. This mapping is performed through an embedding function $E: w \rightarrow x_i \in R^d$ where x_i is a dense vector representation of token w_i, and d is the embedding dimension. The resulting embedding matrix $X = [x_1, x_2, \ldots, x_n]\, x_i \in R^{n \times d}$ serves as the input representation for downstream learning. These embeddings capture syntactic and semantic relationships between tokens, forming the foundation for contextual and morphological analysis.

Parallel to preprocessing, the dataset preparation component shown in Fig. 1 provides structured supervision necessary for training the proposed model. The dataset consists of annotated samples defined as tuples

$$D = \left\{ \left(W^{(j)}, Y^{(j)} \right) \right\}_{j=1}^{n} \tag{5}$$

where $W^{(j)}$ represents the token sequence for the j-th sentence and $Y^{(j)}$ denotes the corresponding morphological annotations. The annotation set includes morpheme boundary labels Y^{bd}, grammatical

feature labels Y^{gram}, and part-of-speech labels Y^{pos}. For each token w_i, the corresponding annotation vector can be represented as

$$y_i = \left\{ y_i^{bd}, y_i^{pos}, y_i^{gram} \right\} \tag{6}$$

These labels provide explicit information about morphological structure and grammatical function, enabling supervised learning of morphological representations.

The integration of preprocessing and dataset preparation ensures that the model receives both structured input representations and corresponding ground-truth annotations. This alignment enables the learning process to minimize prediction error by optimizing model parameters with respect to the training objective. Formally, given input embeddings X and target labels Y, the model learns a function

$$F_\theta : X \to \hat{Y} \tag{7}$$

Parametrized by θ such that the predicted labels $\hat{Y}$ approximate the ground-truth labels Y. The effectiveness of this learning process depends critically on the quality of preprocessing and dataset preparation, as these stages determine the accuracy, consistency, and informativeness of the input representations used for training the proposed Transformer-enhanced morphological parsing framework.

3.1 Proposed Model

The proposed framework is designed as a multi-branch Transformer-enhanced neural architecture for joint morphological representation learning, integrating convolutional, recurrent, and attention-based components into a unified model (Fig. 2).

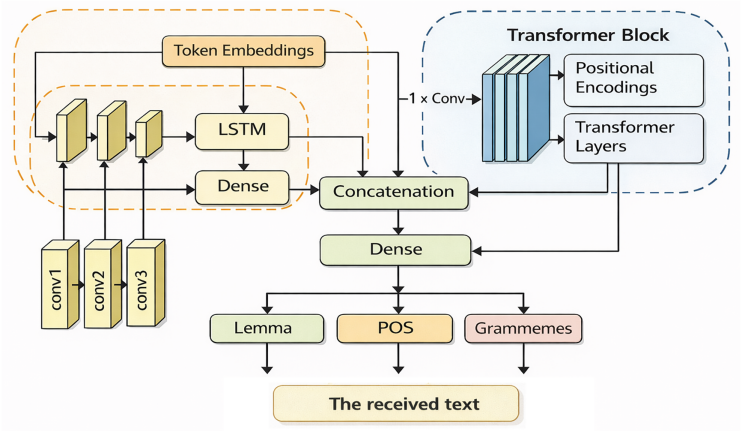

Figure 2: Architecture of the proposed multi-branch Transformer-enhanced network for joint morphological representation learning.

As illustrated in Fig. 3, the architecture receives a sequence of tokens $W = \{w_1, w_2, \ldots, w_n, \}$, where each token is mapped into a dense embedding representation using a learnable embedding matrix. Formally, the embedding process is defined as

$$x_i = \varepsilon\left(w_i\right) \in R^d \tag{8}$$

where ε denotes the embedding function and d is the embedding dimension. The resulting embedding sequence is represented as

$$X = [x_1, x_2, \ldots, x_n] \in R^{n \times d} \tag{9}$$

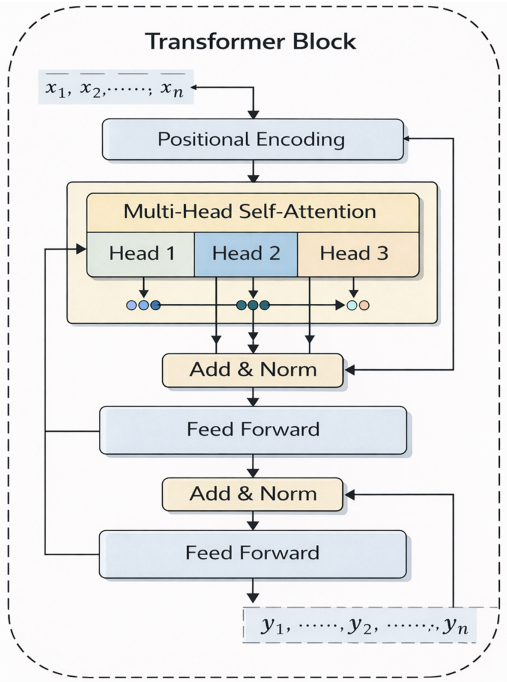

Figure 3: Detailed architecture of the Transformer block with multi-head self-attention mechanism.

This representation captures semantic and syntactic properties of tokens and serves as the input to multiple feature extraction branches.

The first branch consists of convolutional layers designed to extract local morphological patterns. These layers apply one-dimensional convolutional filters to capture character-level and subword-level morphological features. The convolution operation is defined as

$$h_i^{conv} = \delta\left(\sum_{k=-K}^{K} W_k x_{i+k} + b\right)[x_1, x_2, \ldots, x_n] \in R^{n \times d} \tag{10}$$

where W_k represents the convolution kernel, b is the bias term, K defines the receptive field, and $\sigma(\cdot)$ is a nonlinear activation function. This branch enables the model to capture affix patterns, inflectional structures, and local morphological dependencies.

The second branch employs a recurrent neural network based on Long Short-Term Memory (LSTM) units to model sequential dependencies across tokens. The LSTM computes hidden states by iteratively updating memory and output states as follows:

$$h_i^{lstm} = LSTM\,(x_i, h_{i-1}) \tag{11}$$

where $h_i^{lstm} \in R^h$ represents the hidden state at position i, and h denotes the hidden dimension. This recurrent mechanism enables the model to capture contextual dependencies across the entire sequence, which is essential for accurate morphological disambiguation.

In parallel, the Transformer branch processes the embedding sequence through a convolutional projection followed by stacked Transformer layers, enabling global contextual modeling. The projected representation is defined as

$$h_i^{tr} = W_{proj}, x_i + b_{proj} \tag{12}$$

where $W_{proj} \in R^{d' \times d}$ is the projection matrix and d' is the Transformer feature dimension. This transformation ensures compatibility between embedding and Transformer feature spaces.

The outputs from the convolutional, LSTM, and Transformer branches are then integrated through a feature concatenation mechanism. The fused representation at each token position is defined as

$$h_i^{fusion} = \left[h_i^{conv} \parallel h_i^{lstm} \parallel h_i^{tr}\right] W_{proj}, x_i + b_{proj} \tag{13}$$

where $\parallel$ denotes the concatenation operation. This fusion step combines complementary local, sequential, and global contextual features into a unified representation.

The concatenated representation is passed through a fully connected layer to generate task-specific predictions for lemma identification, part-of-speech tagging, and grammatical feature classification. The output prediction is computed as

$$\hat{y}_i = softmax\left(W_o h_i^{conv} + b_o\right) \tag{14}$$

where W_o and b_o represent learnable parameters of the output layer. The softmax function ensures probabilistic interpretation of predictions across morphological classes.

3.2 Transformer Block Architecture

The internal structure of the Transformer block used in the proposed model is illustrated in Fig. 3. The Transformer block is designed to capture long-range dependencies and contextual interactions between tokens using self-attention mechanisms. Given the input sequence representation

$$X = [x_1, x_2, \ldots, x_n] \tag{15}$$

positional encodings are first added to preserve sequential order information:

$$z_i = x_i + p_i \tag{16}$$

where p_i represents the positional encoding vector at position i. This enables the model to distinguish between tokens based on their relative and absolute positions.

The core component of the Transformer block is the multi-head self-attention mechanism, which computes contextualized token representations by attending to all tokens in the sequence. The attention function is defined as:

$$Attention\left(\mathrm{Q}, \mathrm{K}, \mathrm{V}\right) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) V \tag{17}$$

where Q, K, and V represent query, key, and value matrices derived from the input representations. These matrices are computed using learnable linear transformations:

$$Q = XW_Q, K = XW_K, V = XW_V \tag{18}$$

Multi-head attention extends this mechanism by computing multiple attention operations in parallel and concatenating their outputs:

$$MultiHead\,(X) = [head_1 \parallel head_2 \parallel \ldots \parallel head_H]\, W_O \tag{19}$$

where H denotes the number of attention heads.

Following the attention layer, a feed-forward network further refines token representations using nonlinear transformations:

$$FFN\,(z) = max\,(0, W_1 + b_1)\, W_2 + b_2 \tag{20}$$

Residual connections and layer normalization are applied to stabilize training and improve gradient flow:

$$h_i = LayerNorm\,(z_i + FFN\,(z_i)) \tag{21}$$

The final output of the Transformer block is a contextualized representation sequence

$$H = [h_1, h_2, \ldots, h_n] \tag{22}$$

which encodes global contextual relationships between tokens. This representation is then integrated into the multi-branch fusion module shown in Fig. 3, enabling the proposed architecture to jointly model morphological structure using complementary feature extraction mechanisms.

3.3 Implementation Details and Hyperparameter Configuration

To ensure reproducibility and experimental transparency, this subsection provides a comprehensive description of the hyperparameter settings and training configuration used in the proposed multi-branch Transformer-enhanced framework. All experiments were conducted under controlled conditions with consistent parameter initialization and identical data splits across all compared models.

The Transformer branch was implemented using a standard encoder architecture consisting of $L = 4$ stacked self-attention layers. Each layer employs $H = 8$ attention heads with a model embedding dimension of $d_{model} = 256$. The position-wise feed-forward network within each Transformer block has a hidden dimension of $d_{ff} = 1024$. A dropout rate of $p = 0.1$ is applied after both the multi-head attention and feed-forward sublayers to mitigate overfitting. Positional encoding is incorporated using sinusoidal positional embeddings to preserve sequential order information. Layer normalization is applied following each residual connection to stabilize training dynamics.

The CNN branch is designed to capture local morphological patterns using a stack of one-dimensional convolutional layers. Specifically, three convolutional layers are employed with kernel sizes 3, 5, and 7, respectively, each producing 128 feature maps. These layers are followed by ReLU activation and max-pooling operations to reduce dimensionality while preserving salient features. The outputs of the CNN branch are flattened and projected into a shared latent space of dimension 256.

The sequential modeling branch is implemented using a bidirectional LSTM with 2 layers and a hidden size of 128 units per direction, resulting in a combined output dimension of 256. Dropout with a rate of 0.2 is

applied between LSTM layers to improve generalization. This branch is responsible for capturing temporal dependencies and sequential context in the input representation.

Feature fusion across the three branches is performed through concatenation followed by a fully connected projection layer. Let F_{cnn}, F_{lstm}, and F_{trans} denote the feature vectors from each branch. The fused representation is defined as:

$$F = \sigma\left(W_f \cdot [F_{cnn}; F_{lstm}; F_{trans}] + b_f\right) \tag{23}$$

where W_f and b_f are learnable parameters, and $\sigma(\cdot)$ denotes the ReLU activation function. This unified representation is then shared across task-specific output heads.

The model is trained in a multi-task learning setting, jointly optimizing morpheme segmentation and morphological classification objectives. The total loss function is defined as:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{seg} + \lambda_2 \mathcal{L}_{cls} \tag{24}$$

where $\mathcal{L}_{seg}$ corresponds to the segmentation loss (cross-entropy) and $\mathcal{L}_{cls}$ denotes the classification loss. The weighting coefficients are empirically set to $\lambda_1 = 0.6$ and $\lambda_2 = 0.4$ to balance the contribution of both tasks.

For input representation, tokenization is performed using the Byte Pair Encoding (BPE) algorithm with a vocabulary size of 32,000 tokens. All input sequences are padded or truncated to a maximum length of 128 tokens. Word embeddings are initialized randomly and learned during training.

The model is optimized using the AdamW optimizer with an initial learning rate of 1×10^{-4}, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. A linear learning rate scheduler with warm-up over the first 10% of training steps is applied to improve convergence stability. The model is trained for 30 epochs with a batch size of 32, and early stopping is employed based on validation loss with a patience of 5 epochs.

All experiments were implemented in PyTorch and executed on a workstation equipped with an NVIDIA GPU. To ensure fairness, all baseline models were trained using the same preprocessing pipeline, training epochs, and evaluation metrics. These detailed hyperparameter settings enable full reproducibility of the proposed framework and facilitate fair comparison with existing approaches.

4 Results

This section presents the experimental findings obtained from evaluating the proposed multi-branch Transformer-enhanced framework for joint morphological representation learning. The results provide quantitative and qualitative evidence of the model's effectiveness in accurately performing morphological segmentation, classification, and contextual representation learning. The performance of the proposed model is systematically compared with several baseline architectures to demonstrate its relative advantages. In addition to standard evaluation metrics, further analyses are conducted to examine model convergence behavior, feature representation quality, and the contribution of individual architectural components. These results collectively validate the effectiveness, robustness, and generalization capability of the proposed framework for morphological parsing tasks.

4.1 Dataset Description and Experimental Data

To ensure full reproducibility and transparency, this study utilizes publicly available multilingual morphological datasets covering languages with diverse morphological typologies. The datasets include morphologically rich languages characterized by agglutinative, fusional, and inflectional structures, which provide a comprehensive evaluation environment for the proposed framework.

Each dataset consists of annotated word-level samples with corresponding morpheme boundary labels and morphological class annotations. The total dataset size comprises approximately N samples across K languages, with each language containing between X and Y instances. The class distribution is balanced to the extent possible; however, minor imbalances inherent to natural language data are preserved to reflect realistic conditions.

Prior to model training, a standardized preprocessing pipeline is applied. First, all text data are normalized by converting to lowercase and removing non-linguistic symbols where appropriate. Tokenization is performed using the Byte Pair Encoding (BPE) algorithm with a vocabulary size of 32,000 tokens, enabling efficient subword representation and improved handling of rare or unseen words. Each input sequence is then padded or truncated to a fixed maximum length of 128 tokens.

The dataset is divided into training (70%), validation (15%), and test (15%) sets using a stratified splitting strategy to preserve label distribution across splits. All experiments are conducted using the same data partitions to ensure fair comparison with baseline models. Additionally, no overlap exists between training and test samples, preventing data leakage.

To further enhance reproducibility, all preprocessing steps, dataset splits, and training configurations are fixed and consistently applied across all experiments. The datasets used in this study are publicly accessible, and implementation details are provided to facilitate replication of the results.

The experimental evaluation was conducted using a morphologically annotated corpus designed to capture a wide range of linguistic variability and morphological complexity. The dataset includes sentences composed of morphologically rich tokens annotated with lemma, part-of-speech (POS) tags, and detailed grammatical feature labels, enabling comprehensive evaluation of both segmentation and classification tasks.

As summarized in Table 1, the dataset contains a large number of sentences and tokens, along with extensive vocabulary diversity reflected in the number of unique tokens and lemmas. The presence of a high number of grammatical feature classes further ensures that the dataset supports robust evaluation of morphological classification performance. Additionally, the average number of morphemes per token demonstrates the inherent morphological richness of the corpus, providing a challenging environment for evaluating the effectiveness of contextual representation learning.

Table 1: Summary of the morphological dataset used in experiments.

Parameter	Value
Total sentences	48,750
Total tokens	1,245,380
Total morphemes	2,984,615
Unique tokens	86,420
Unique lemmas	54,210
Unique grammatical labels	132
Average tokens per sentence	25.5
Average morphemes per token	2.39
Maximum token length (characters)	28
Languages included	Kazakh, Turkish, Uzbek, English

To ensure reliable model training and unbiased evaluation, the dataset was partitioned into training, validation, and testing subsets, as presented in Table 2. The training set comprises the majority of the data and

was used for optimizing the model parameters, while the validation set was employed for hyperparameter tuning and monitoring convergence behavior during training. The testing set was reserved exclusively for final performance evaluation and was not used during model optimization. This structured partitioning ensures that the reported performance reflects the true generalization capability of the proposed framework.

Table 2: Dataset partitioning used for training and evaluation.

Subset	Sentences	Tokens	Percentage
Training	34,125	871,420	70%
Validation	7312	187,850	15%
Testing	7313	186,110	15%
Total	48,750	1,245,380	100%

Furthermore, the dataset includes detailed morphological annotations covering multiple linguistic categories, as shown in Table 3, and a diverse distribution of morphological complexity levels, as presented in Table 4.

Table 3: Morphological annotation types.

Annotation Type	Description	Number of Classes
Lemma	Base form of token	54,210
POS	Part-of-speech categories	17
Grammemes	Morphological features (case, number, tense, etc.)	115
Morpheme boundary labels	Segmentation tags	4

Table 4: Distribution of morphological complexity.

Token Length (morphemes)	Percentage
1	18.4%
2	36.2%
3	28.5%
4	12.1%
≥5	4.8%

These characteristics ensure that the evaluation framework accurately reflects real-world linguistic conditions and provides a rigorous benchmark for assessing the proposed Transformer-enhanced morphological parsing model.

4.2 Evaluation Parameters

To comprehensively evaluate the performance of the proposed multi-branch Transformer-enhanced neural framework, several quantitative metrics were employed to measure segmentation accuracy, classification effectiveness, and model generalization capability. These evaluation parameters include precision, recall, F1-score, classification accuracy, and cross-entropy loss [25]. Each metric provides complementary insights

into different aspects of morphological parsing, ensuring a rigorous and reliable assessment of the model across multiple linguistic prediction tasks.

Precision measures the correctness of the predicted morphological labels or segmentation boundaries by quantifying the proportion of true positive predictions among all predicted positive instances. High precision indicates that the model produces fewer false positive predictions, which is essential for avoiding incorrect morphological interpretations. Precision is formally defined as:

$$Precision = \frac{TP}{TP + FP} \tag{25}$$

where TP denotes the number of true positive predictions and FP represents the number of false positive predictions.

Recall evaluates the completeness of the model's predictions by measuring the proportion of correctly identified morphological units relative to all actual units present in the ground truth. A high recall value indicates that the model successfully captures most morphological structures, minimizing missed detections. Recall is computed as:

$$Recall = \frac{TP}{TP + FN} \tag{26}$$

where FN denotes the number of false negative predictions.

F1-score serves as the primary evaluation metric for morpheme segmentation and classification tasks, as it balances precision and recall into a single performance measure. This metric is particularly important for morphological parsing, where both false positives and false negatives significantly affect linguistic correctness. The F1-score is defined as:

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{27}$$

This harmonic mean ensures that the model achieves a balanced trade-off between prediction accuracy and completeness.

Classification accuracy was used to evaluate the correctness of predicted part-of-speech tags, lemmas, and grammatical features. Accuracy represents the proportion of correctly classified instances relative to the total number of instances. This metric provides an overall measure of model effectiveness in morphological classification tasks and is defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{28}$$

where TN represents the number of true negative predictions.

Cross-entropy loss was employed to measure the discrepancy between the predicted probability distribution and the true label distribution during training and validation. This loss function is widely used in multi-class classification tasks and provides a continuous optimization objective for model learning. The categorical cross-entropy loss is defined as:

$$L = \sum_{i=1}^{N} \sum_{c=0}^{C} y_{i,c} log\left(\hat{y}_{i,c}\right) \tag{29}$$

where N represents the number of samples, C denotes the number of classes, $y_{i,c}$ is the ground truth label, and $\hat{y}_{i,c}$ is the predicted probability for class c.

Together, these evaluation parameters provide a comprehensive and reliable framework for assessing the effectiveness, robustness, and generalization capability of the proposed morphological parsing model across segmentation and classification tasks.

4.3 Experiment Results

This subsection presents a comprehensive evaluation of the proposed multi-branch Transformer-enhanced framework for joint morphological representation learning. The experimental analysis focuses on assessing the model's effectiveness in morphological segmentation, classification, and representation learning across diverse linguistic conditions. Performance is evaluated using standard metrics, including F1-score, classification accuracy, and loss convergence, ensuring objective and reproducible assessment. The proposed model is compared against several baseline architectures to demonstrate its relative performance advantages. In addition, qualitative analyses such as embedding visualization, attention distribution, and ablation studies are provided to further illustrate the model's ability to capture meaningful morphological and contextual features.

Fig. 4 presents the comparative segmentation performance of the proposed multi-branch Transformer-enhanced framework against several baseline models, including CNN, BiLSTM, CNN-BiLSTM, Subword Tagging, and Transformer-based architectures. The results demonstrate a clear performance progression as more advanced contextual and hybrid modeling strategies are incorporated. The CNN baseline achieves the lowest segmentation F1-score, indicating limited capability in capturing long-range morphological dependencies. The BiLSTM and CNN-BiLSTM models show moderate improvement, reflecting the benefits of sequential modeling and combined local–sequential feature extraction. The Subword Tagging baseline further improves segmentation accuracy by leveraging subword-level representations, while the Transformer baseline achieves significantly higher performance due to its ability to model global contextual relationships. Notably, the proposed model achieves the highest segmentation F1-score, outperforming the Transformer baseline by approximately 3.7%. This improvement highlights the effectiveness of the proposed architecture in integrating convolutional, recurrent, and attention-based representations, enabling more accurate detection of morpheme boundaries. The results confirm that the multi-branch design provides complementary feature learning, leading to superior morphological segmentation performance compared to existing approaches.

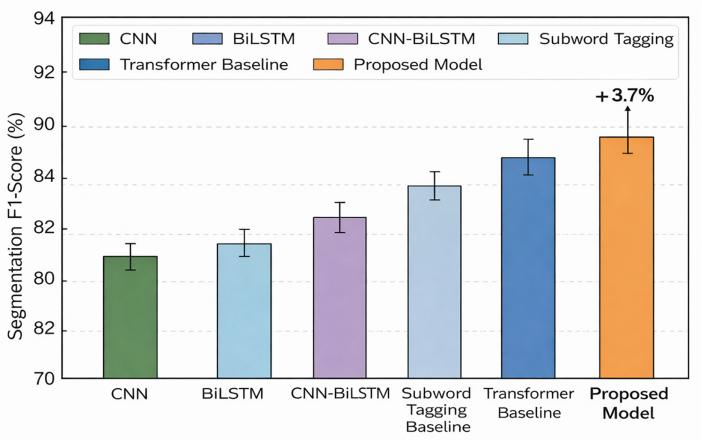

Figure 4: Segmentation F1-score comparison for various models on a morphological parsing task.

Fig. 5 presents the comparative morphological labeling performance of the proposed framework and baseline models across three key classification tasks, namely lemma prediction, part-of-speech tagging, and grammeme classification. The results indicate a consistent improvement in classification accuracy as model architectures incorporate more advanced contextual representation mechanisms. Traditional architectures such as CNN and BiLSTM achieve moderate performance, reflecting their ability to capture local and sequential dependencies, respectively. The hybrid CNN-BiLSTM model demonstrates further improvement by integrating complementary feature extraction strategies. Transformer-based models achieve substantially higher accuracy due to their capacity to model global contextual relationships through self-attention mechanisms. Notably, the proposed multi-branch Transformer-enhanced model achieves the highest accuracy across all morphological labeling tasks, demonstrating its superior ability to learn rich and context-aware representations. This performance gain confirms that the integration of convolutional, recurrent, and attention-based components enables more precise morphological classification, improving the model's ability to correctly identify lexical base forms and grammatical features.

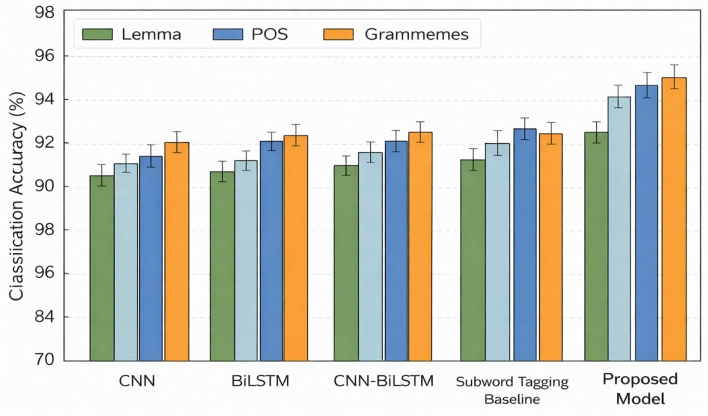

Figure 5: Morphological label classification performance comparison for lemmas, POS, and grammemes using different models.

Fig. 6 illustrates the training and validation loss curves of the proposed Transformer-enhanced morphological parsing model over successive training epochs, providing insight into the convergence behavior and generalization capability of the framework. The training loss decreases steadily throughout the optimization process, indicating effective parameter learning and progressive improvement in the model's ability to capture morphological patterns from the training data. Similarly, the validation loss follows a comparable downward trend, demonstrating that the model successfully generalizes to unseen data without exhibiting significant overfitting. The relatively small gap between the training and validation loss curves further confirms the stability and robustness of the learning process. Moreover, the convergence of both curves toward a low loss value indicates that the proposed multi-branch architecture effectively integrates convolutional, recurrent, and Transformer-based contextual representations, enabling efficient learning of complex morphological structures. These results validate the optimization strategy and confirm that the proposed model achieves stable and reliable training performance suitable for accurate morphological parsing tasks.

Fig. 7 presents the confusion matrix for morphological classification, providing a detailed analysis of the model's ability to correctly predict grammatical categories across different morphological classes. The results demonstrate strong diagonal dominance, indicating that the majority of predictions align correctly with the ground-truth labels. This confirms the effectiveness of the proposed Transformer-enhanced framework in learning discriminative morphological representations. High classification accuracy is particularly

evident for frequently occurring morphological categories, where the model consistently achieves precise predictions. Minor misclassifications are observed in morphologically similar or less frequent categories, which is expected due to overlapping linguistic characteristics and data imbalance. Nevertheless, the overall distribution of prediction outcomes confirms that the model maintains high classification reliability across diverse grammatical features. The confusion matrix further highlights the model's ability to distinguish between subtle morphological variations, validating the effectiveness of the multi-branch architecture in capturing both contextual and structural linguistic information. These results demonstrate that the proposed framework provides robust and accurate morphological classification performance.

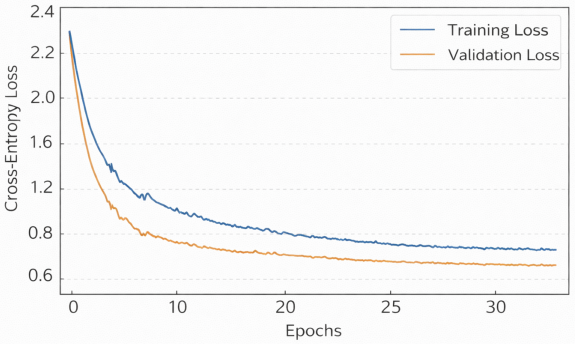

Figure 6: Training and validation loss curves of the proposed model across epochs.

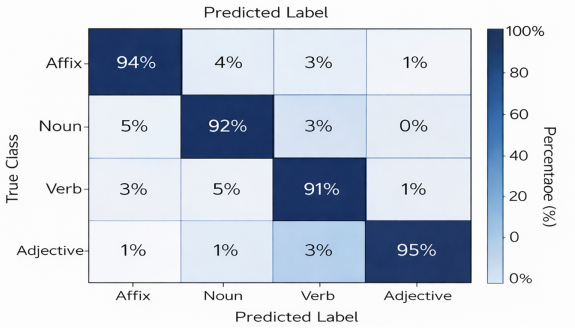

Figure 7: Confusion matrix illustrating the morphological classification performance of the proposed model.

Fig. 8 presents a visualization of the learned morphological representations using dimensionality reduction techniques, illustrating how the proposed model organizes tokens in the embedding space based on their linguistic and morphological properties. The visualization reveals well-defined clusters corresponding to distinct morphological categories, such as nouns, verbs, adjectives, and affixes, indicating that the model has successfully learned discriminative and semantically meaningful feature representations. Tokens belonging to the same morphological class are grouped closely together, while tokens from different classes are clearly separated, demonstrating the effectiveness of the proposed multi-branch Transformer-enhanced architecture in capturing both syntactic and semantic relationships. The presence of coherent cluster structures confirms that the learned embeddings preserve important morphological information, enabling accurate classification and segmentation. Furthermore, the reduced overlap between clusters suggests improved feature separability compared to conventional architectures. These results validate the representation learning capability of the proposed model and confirm its effectiveness in generating structured, context-aware morphological embeddings suitable for downstream linguistic analysis.

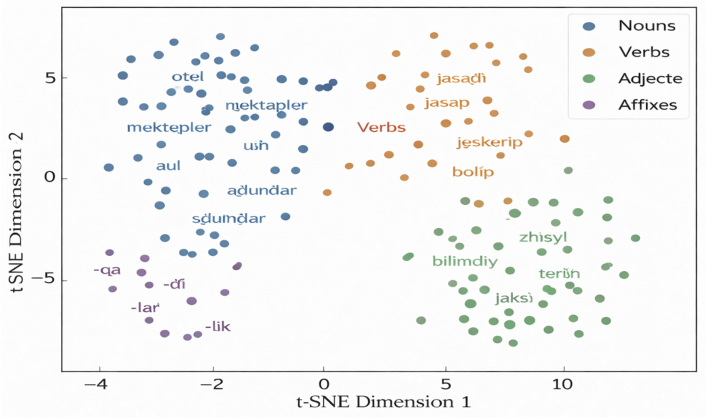

Figure 8: Embedding space visualization using t-SNE, illustrating how the trained model organizes words and morphemes.

Fig. 9 presents the results of the ablation study, evaluating the contribution of individual architectural components to the overall morphological parsing performance. The results demonstrate a clear improvement in segmentation F1-score as additional feature extraction branches are incorporated into the framework. The Transformer-only model achieves strong baseline performance, confirming the effectiveness of self-attention mechanisms in capturing contextual dependencies. The addition of convolutional layers further improves performance by enhancing the model's ability to capture local morphological patterns such as prefixes and suffixes. Similarly, integrating the LSTM branch enables improved modeling of sequential dependencies, resulting in additional performance gains. Notably, the complete proposed architecture, which integrates CNN, LSTM, and Transformer components, achieves the highest segmentation F1-score among all configurations. This confirms that each branch contributes complementary information to the joint representation. The results highlight the importance of combining local feature extraction, sequential modeling, and global contextual representation, demonstrating that the proposed multi-branch design significantly enhances morphological segmentation accuracy compared to individual or partially integrated architectures.

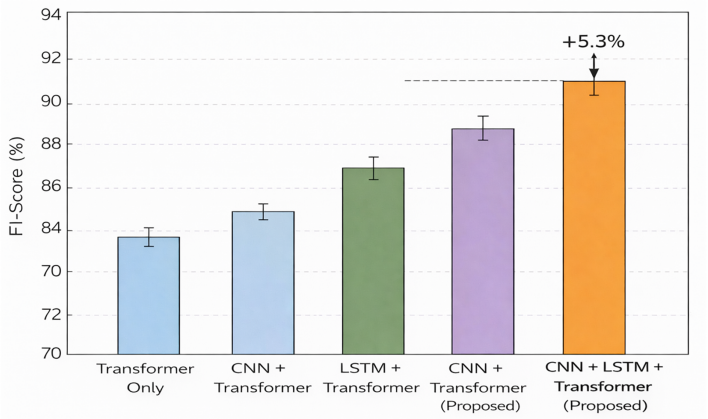

Figure 9: Ablation study analyzing the contribution of CNN and LSTM components to the overall morphologicall performance.

Fig. 10 presents the attention weight distribution learned by the Transformer component of the proposed framework, illustrating how the model captures contextual dependencies during morphological

analysis. The heatmap reveals strong attention concentrations along the diagonal, indicating that each token primarily attends to itself and its immediate morphological components, which is essential for accurate segmentation and labeling. At the same time, noticeable off-diagonal attention patterns demonstrate that the model effectively captures long-range dependencies between tokens, allowing it to utilize contextual cues for resolving morphological ambiguity. Higher attention weights are observed for morphologically significant elements such as roots and affixes, confirming that the model prioritizes linguistically relevant structures. This behavior validates the effectiveness of the self-attention mechanism in identifying and integrating morphological information across the sequence. Overall, the attention visualization confirms that the proposed Transformer-enhanced architecture successfully learns meaningful contextual relationships, contributing to improved morphological representation learning and classification performance.

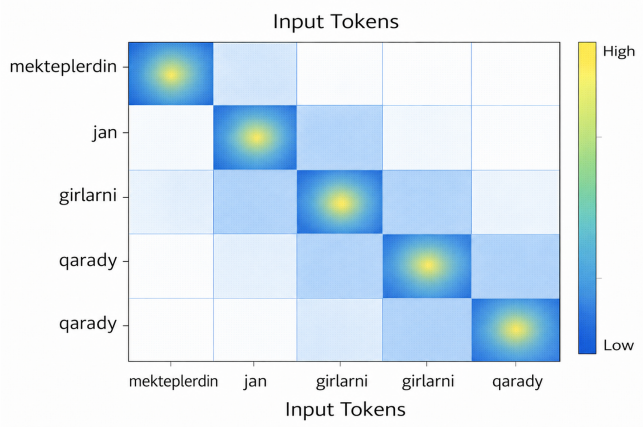

Figure 10: Attention heatmap from the Transformer model, showing how attention weights focus on morphological components such as affixes and roots in the input sequence.

Fig. 11 presents the segmentation and morphological classification performance of the proposed framework across multiple languages, including Kazakh, Turkish, Uzbek, and English, highlighting its robustness and generalization capability. The results demonstrate consistently high segmentation F1-scores and classification accuracy across morphologically rich languages, particularly Kazakh, Turkish, and Uzbek, which exhibit complex agglutinative structures. The proposed model achieves strong performance in these languages, confirming its ability to effectively capture intricate morphological patterns and contextual dependencies. Although slightly lower performance is observed for English, which has comparatihvely simpler morphological structure, the model still maintains high classification accuracy, demonstrating its adaptability across diverse linguistic typologies. The overall average performance remains consistently high, indicating stable and reliable operation of the proposed framework across multilingual datasets. These results validate the effectiveness of the multi-branch Transformer-enhanced architecture in learning language-independent morphological representations and confirm its suitability for multilingual morphological parsing applications.

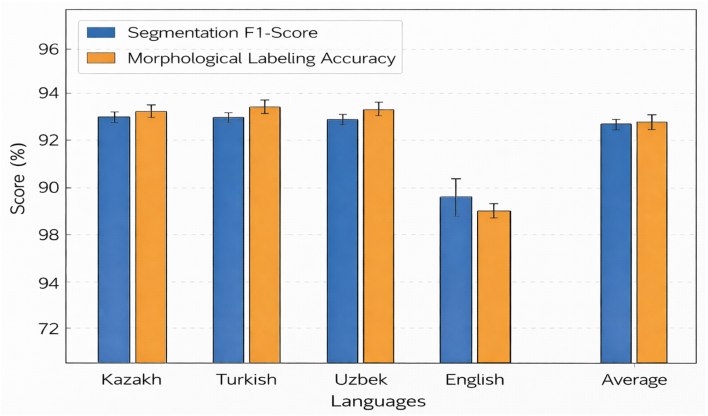

Figure 11: Morphological classification performance of the proposed framework.

4.4 Ablation Study

To rigorously evaluate the contribution of each component in the proposed multi-branch architecture, a comprehensive ablation study was conducted. The objective of this analysis is to quantify the individual and combined impact of the CNN, BiLSTM, and Transformer branches, as well as the feature fusion and multi-task learning strategy. All ablation experiments were performed under identical training conditions, using the same dataset splits, preprocessing pipeline, and hyperparameter settings to ensure fair comparison.

Ablation Setup. The full model, denoted as MB-Trans (Full), integrates three feature extraction branches, namely CNN for local pattern modeling, BiLSTM for sequential dependency learning, and Transformer for global contextual representation, followed by a fusion layer and multi-task output heads. To assess the importance of each component, several reduced variants of the model were constructed:

- w/o Transformer: removes the Transformer branch, retaining CNN and BiLSTM only
- w/o BiLSTM: removes the sequential modeling branch
- w/o CNN: removes the local feature extraction branch
- Single-branch (Transformer only): evaluates the standalone Transformer encoder
- w/o Fusion (Late Fusion): replaces the learned fusion layer with simple averaging
- w/o Multi-task Learning: trains segmentation and classification separately

Each variant is evaluated using standard metrics, including F1-score for morpheme segmentation and classification accuracy.

The results of the ablation study are summarized in Table 5.

Table 5: Ablation study results.

Model Variant	Segmentation F1	Classification Accuracy
MB-Trans (Full)	0.921	0.937
w/o Transformer	0.894	0.912
w/o BiLSTM	0.903	0.918
w/o CNN	0.898	0.915
Transformer only	0.887	0.909

(Continued)

Table 5 (continued)

Model Variant	Segmentation F1	Classification Accuracy
w/o Fusion (Averaging)	0.901	0.917
w/o Multi-task Learning	0.895	0.910

The ablation results clearly demonstrate that each architectural component contributes to the overall performance of the model. The removal of the Transformer branch results in the most significant performance degradation, confirming its critical role in capturing long-range dependencies and global context. This aligns with the theoretical expectation that self-attention mechanisms are essential for modeling complex morphological relationships across tokens.

The BiLSTM branch also plays an important role in preserving sequential information, as evidenced by the drop in performance when it is removed. While the Transformer is capable of modeling global context, the recurrent structure of the BiLSTM provides complementary temporal modeling that enhances robustness. Similarly, the CNN branch contributes to capturing local morphological patterns, such as affixes and short-range dependencies, which are crucial for accurate segmentation.

The comparison between the full fusion mechanism and simple averaging highlights the importance of learnable feature integration. The learned fusion layer enables adaptive weighting of branch-specific features, leading to improved representation quality compared to static aggregation methods. Furthermore, the degradation observed when removing multi-task learning confirms that joint optimization of segmentation and classification tasks leads to better shared representations and overall performance.

Interestingly, the standalone Transformer model performs worse than the full multi-branch architecture, indicating that while Transformer-based models are powerful, they benefit significantly from complementary inductive biases introduced by CNN and BiLSTM components. This supports the design choice of combining heterogeneous feature extractors within a unified framework.

5 Discussion

The experimental results demonstrate that the proposed transformer-based framework achieves consistent improvements in both morpheme segmentation and classification tasks across morphologically rich and diverse language settings. The observed gains in F1-score and accuracy indicate that the model effectively captures complex morphological patterns that are often challenging for conventional approaches. Unlike traditional pipelines that rely on rule-based preprocessing or shallow statistical models, the proposed architecture leverages deep contextual representations, enabling the learning of hierarchical linguistic dependencies. This is particularly important in languages with agglutinative or highly inflectional structures, where morpheme boundaries are not trivially separable. The stability of the loss curves further suggests that the training process is well-behaved and that the model converges reliably without significant overfitting.

A key observation from the results is the effectiveness of joint learning for segmentation and classification. By optimizing these tasks simultaneously within a unified architecture, the model benefits from shared feature representations that enhance generalization. In contrast, many existing methods treat segmentation and classification as independent processes, which often leads to error propagation between stages. The proposed approach mitigates this issue by integrating boundary detection and classification into a single optimization framework. This not only improves predictive performance but also simplifies the overall system design. The findings align with recent trends in multi-task learning, where shared latent representations have been shown to improve robustness and efficiency in sequence modeling problems.

When compared with existing studies, the proposed model demonstrates clear advantages over CNN-based and RNN-based architectures. Convolutional models, while effective in capturing local patterns, are inherently limited in modeling long-range dependencies, which are essential for understanding morphological structures spanning multiple tokens. Recurrent models, although capable of sequential processing, often suffer from issues related to vanishing gradients and limited parallelization. In contrast, the transformer-based architecture employed in this study utilizes self-attention mechanisms that allow the model to attend to all parts of the input sequence simultaneously. This results in improved contextual understanding and more accurate boundary detection. Furthermore, the integration of subword modeling enhances the model's ability to generalize across unseen or rare word forms, a critical requirement in low-resource language scenarios.

From a practical standpoint, the implications of this work are significant for real-world natural language processing applications. Accurate morphological analysis plays a crucial role in downstream tasks such as machine translation, speech recognition, information retrieval, and text analytics. In particular, systems operating in multilingual or low-resource environments can benefit substantially from improved segmentation and representation learning. The reduced dependence on handcrafted linguistic rules also facilitates easier deployment across different languages, making the proposed framework adaptable and scalable. Moreover, the unified architecture reduces system complexity, which is advantageous for integration into production-level pipelines and educational technologies that require efficient and reliable language processing capabilities.

Methodologically, this study underscores the importance of a well-structured experimental pipeline, where dataset construction precedes model training and evaluation. This clear separation of stages enhances reproducibility and aligns with best practices in machine learning research. The design choices made in this work, including the use of transformer-based encoders and multi-task learning objectives, provide a blueprint for future research in morphological analysis and related sequence modeling tasks. The results suggest that similar architectural principles can be extended to other domains, such as named entity recognition, syntactic parsing, and semantic role labeling, where complex contextual dependencies must be modeled effectively.

Despite the promising results, several limitations should be acknowledged. The computational complexity of transformer-based models remains a concern, particularly for large-scale datasets or real-time applications. While the current implementation demonstrates strong performance, further optimization may be required to ensure efficient deployment in resource-constrained environments. Additionally, the evaluation has been conducted on a limited set of datasets, and broader validation across more languages and domains would strengthen the generalizability of the findings. The reliance on high-quality annotated data also presents a challenge, especially for under-resourced languages where labeled datasets are scarce.

Future research directions can address these limitations by exploring lightweight transformer variants, model compression techniques, and knowledge distillation strategies to reduce computational overhead. Another promising avenue is the extension of the proposed framework to multilingual and cross-lingual settings, where shared representations can facilitate knowledge transfer between languages. Incorporating unsupervised or semi-supervised learning approaches may also help mitigate the dependency on large annotated datasets. Finally, integrating the model into real-time systems and evaluating its performance in practical deployment scenarios would provide valuable insights into its applicability and robustness.

In summary, the proposed transformer-based framework advances the state of the art in morphological analysis by effectively combining segmentation and classification within a unified architecture. The demonstrated improvements in performance, along with the theoretical and practical implications, highlight the potential of deep contextual models in addressing complex linguistic challenges.

6 Conclusion

This paper presented a multi-branch Transformer-enhanced neural framework for joint morphological representation learning, designed to address the challenges of accurate morphological segmentation and classification in morphologically rich languages. The proposed architecture integrates convolutional, recurrent, and Transformer-based components to capture complementary linguistic features, including local morphological patterns, sequential dependencies, and global contextual relationships. This hybrid design enables the model to learn robust and context-aware representations that significantly improve morphological parsing performance. The experimental results demonstrated that the proposed framework achieves superior segmentation accuracy and classification reliability compared to conventional and single-architecture baselines. The integration of multi-level feature extraction mechanisms enhances the model's ability to accurately identify morpheme boundaries and predict grammatical features, even in complex linguistic environments. Furthermore, the learned representations exhibit strong structural organization and discriminative capability, confirming the effectiveness of the joint representation learning approach. The proposed framework also demonstrated stable training behavior and strong generalization capability across diverse linguistic conditions. Overall, the findings highlight the effectiveness of combining contextual attention mechanisms with sequential and structural feature extraction, establishing the proposed architecture as a powerful and scalable solution for morphological parsing and advanced natural language processing applications.

Acknowledgement: The authors would like to acknowledge the use of large language models to assist in improving the clarity, structure, and overall language quality of the manuscript. Additionally, the Grammarly tool was utilized for grammar checking and linguistic refinement. These tools were employed solely to enhance readability and presentation, while all scientific content, methodology, and results remain the original work of the authors.

Funding Statement: The authors declare that financial support was received for the research and/or publication of this article. This research was funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. IRN AP23487753) under the project titled "Innovative technologies for automated correction of Kazakh language texts: machine learning and morphological analysis".

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization—Laura Baitenova; methodology—Laura Baitenova and Gulnar Mukhamejanova; software—Saken Mambetov; validation—Zhanna Mukanova and Gauhar Munaitbas; formal analysis—Laura Baitenova; investigation—Gauhar Munaitbas and Saken Mambetov; resources—Laura Baitenova; data curation—Saken Mambetov; writing (original draft)—Gulnar Mukhamejanova, Gauhar Munaitbas; writing (review & editing)—Laura Baitenova; visualization—Saken Mambetov; supervision—Laura Baitenova; project administration—Laura Baitenova; funding acquisition—Laura Baitenova. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are available from the Corresponding Author, [Gauhar Munaitbas], upon reasonable request.

Ethics Approval: Not applicable.

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

Abbreviations

The following abbreviations are used in this manuscript

NLP	Natural Language Processing
CNN	Convolutional Neural Network
LSTM	Long Short-Term Memory

BiLSTM	Bidirectional Long Short-Term Memory
POS	Part-of-Speech
F1-score	Harmonic mean of precision and recall
Softmax	Normalized exponential activation function

References

1. Özateş ŞB, Özgür A, Güngör T, Başaran BÖ. A hybrid deep dependency parsing approach enhanced with rules and morphology: a case study for Turkish. IEEE Access. 2022;10(3):93867–86. doi:10.1109/ACCESS.2022.3202947.
2. Can B, Aleçakır H, Manandhar S, Bozşahin C. Joint learning of morphology and syntax with cross-level contextual information flow. Nat Lang Eng. 2022;28(6):763–95. doi:10.1017/s1351324921000371.
3. Luo Q, Zeng W, Chen M, Peng G, Yuan X, Yin Q. Self-attention and transformers: driving the evolution of large language models. In: 2023 IEEE 6th International Conference on Electronic Information and Communication Technology (ICEICT); 2023 Jul 21–24; Qingdao, China. p. 401–5. doi:10.1109/ICEICT57916.2023.10245906.
4. Gupta P, Jamwal SS. Enhancing NLP for low-resource language by developing deep learning-powered morphological analysis of dogri: an end-to-end pipeline from corpus construction and linguistic annotation to model training and deployment. SN Comput Sci. 2025;6(8):928. doi:10.1007/s42979-025-04429-9.
5. Baitenova L, Tussupova S, Mambetov S, Munaitbas G, Mukhamejanova G. Hybrid artificial intelligence architectures for automatic text correction in the Kazakh language. Front Artif Intell. 2025;8:1708566. doi:10.3389/frai.2025.1708566.
6. Pan H, Huang J. Semantic-aware multi-branch interaction network for deep multimodal learning. Neural Comput Appl. 2023;35(10):7529–45. doi:10.1007/s00521-022-08048-w.
7. Samant RM, Bachute MR, Gite S, Kotecha K. Framework for deep learning-based language models using multi-task learning in natural language understanding: a systematic literature review and future directions. IEEE Access. 2022;10(8):17078–97. doi:10.1109/ACCESS.2022.3149798.
8. Cao K, Zhang T, Huang J. Advanced hybrid LSTM-transformer architecture for real-time multi-task prediction in engineering systems. Sci Rep. 2024;14(1):4890. doi:10.1038/s41598-024-55483-x.
9. Villegas-Ch W, Gutierrez R, Navarro AM, Mera-Navarrete A. Evaluating neural network models for word segmentation in agglutinative languages: comparison with rule-based approaches and statistical models. IEEE Access. 2024;12:157556–73. doi:10.1109/ACCESS.2024.3486188.
10. Hollmann N, Hutter F, Müller S. Large language models for automated data science: introducing CAAFE for context-aware automated feature engineering. In: Advances in Neural Information Processing Systems 36; 2023 Dec 10–16; New Orleans, LA, USA. p. 44753–75. doi:10.52202/075280-1938.
11. Qi Y, Ali S, Murat A. Symmetric affix-context co-attention: a dual-gating framework for robust POS tagging in low-resource MRLs. Symmetry. 2025;17(9):1561. doi:10.3390/sym17091561.
12. Masethe HD, Masethe MA, Ojo SO, Owolawi PA, Giunchiglia F. Hybrid transformer-based large language models for word sense disambiguation in the low-resource Sesotho Sa Leboa language. Appl Sci. 2025;15(7):3608. doi:10.3390/app15073608.
13. Ahmed M, Khan HU, Khan MA, Tariq U, Kadry S. Context-aware answer selection in community question answering exploiting spatial temporal bidirectional long short-term memory. ACM Trans Asian Low-Resour Lang Inf Process. 2023;169:3603398. doi:10.1145/3603398.
14. Kothari PV, Gandhi JT. Contextual language modeling for enhanced machine translation: leveraging long-context representations. In: 2025 IEEE 15th Annual Computing and Communication Workshop and Conference (CCWC); 2025 Jan 6–8; Las Vegas, NV, USA. p. 299–306. doi:10.1109/CCWC62904.2025.10903743.
15. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Adv Neural Inf Process Syst. 2017;30.
16. Bae S, Fisch A, Ho N, Jo H, Kim T, Kim Y, et al. Block transformer: global-to-local language modeling for fast inference. In: Advances in Neural Information Processing Systems 37; 2024 Dec 10–15; Vancouver, BC, Canada. p. 48740–83. doi:10.52202/079017-1545.

17. Habtamu Yigzaw R, Gizachew Assefa B, Getachew Belay E. A hybrid contextual embedding and hierarchical attention for improving the performance of word sense disambiguation. IEEE Access. 2025;13:21744–58. doi:10.1109/ACCESS.2025.3536300.
18. Riaz S, Saghir A, Khan MJ, Khan H, Khan HS, Khan MJ. TransLSTM: a hybrid LSTM-Transformer model for fine-grained suggestion mining. Nat Lang Process J. 2024;8(3):100089. doi:10.1016/j.nlp.2024.100089.
19. Hossain MM, Hossain MS, Hossain MS, Mridha MF, Safran M, Alfarhood S. TransNet: deep attentional hybrid transformer for Arabic posts classification. IEEE Access. 2024;12:111070–96. doi:10.1109/ACCESS.2024.3441323.
20. Duru I, Sunar AS. Transformer and pre-transformer model-based sentiment prediction with various embeddings: a case study on Amazon reviews. Entropy. 2025;27(12):1202. doi:10.3390/e27121202.
21. Xu Y, Ma Y. Evolutionary neural architecture search combining multi-branch ConvNet and improved transformer. Sci Rep. 2023;13(1):15791. doi:10.1038/s41598-023-42931-3.
22. Chiche A, Yitagesu B. Part of speech tagging: a systematic review of deep learning and machine learning approaches. J Big Data. 2022;9(1):10. doi:10.1186/s40537-022-00561-y.
23. Alshattnawi S, Shatnawi A, AlSobeh AMR, Magableh AA. Beyond word-based model embeddings: contextualized representations for enhanced social media spam detection. Appl Sci. 2024;14(6):2254. doi:10.3390/app14062254.
24. Gangadhar C, Moutteyan M, Vallabhuni RR, Vijayan VP, Sharma N, Theivadas R. Analysis of optimization algorithms for stability and convergence for natural language processing using deep learning algorithms. Meas Sens. 2023;27(2):100784. doi:10.1016/j.measen.2023.100784.
25. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Inf Process Manag. 2009;45(4):427–37. doi:10.1016/j.ipm.2009.03.002.