CSLRConformer: A Data-Centric Conformer Approach for Continuous Arabic Sign Language Recognition on the Isharah Dataset

Fatimah Mohamed Emad Elden

Department of Computer ans Information Sciences,
Faculty of Graduate Studies for Statistical Research, Cairo University

12422024441586@pg.cu.edu.eg

July 25, 2025

Abstract

The field of Continuous Sign Language Recognition (CSLR) presents a substantial technical challenges including fluid inter sign transitions, absence of temporal boundaries, and co-articulation effects. This paper, conducted for the MSLR 2025 Workshop Challenge at ICCV 2025, addresses signer-independent recognition challenges to develop robust CSLR systems that generalize across diverse signers and improve Arabic sign language recognition. A comprehensive data-centric methodology was developed encompassing systematic feature engineering, preprocessing pipeline design, and model architecture optimization on the Isharah dataset. Primary contributions include: (1) systematic feature engineering through EDA, identifying communicative keypoints while reducing body region noise; (2) robust pre-processing with DBSCAN-based outlier filtering and spatial normalization; (3) novel CSLRConformer architecture adapting Conformer's hybrid CNN-Transformer design for sign language, combining convolutional layers for local temporal dependencies with self-attention for global sequence modeling. The methodology achieved competitive performance with 5.60% Word Error Rate on development set and 12.70% on the test set, demonstrating 75.1% relative WER reduction on development set and 53.6% on test set compared to the best-performing baselines from the original Isharah dataset, and establishing that systematic, data-driven feature engineering coupled with hybrid architectural design provides substantial performance gains for CSLR applications.

Keywords: Continuous Sign Language Recognition (CSLR), Hybrid CNN-Transformer, CSLRConformer

1. Introduction

Sign language serves as the primary communication language for over 70 million deaf individuals worldwide [27], yet the scarcity of continuous sign language recognition (CSLR) datasets remains a critical bottleneck for developing new assistive technologies that bridge the communication gap between deaf and hearing communities. While Isolated Sign Language Recognition (ISLR) has achieved considerable progress [21], the field increasingly demands advancement toward Continuous Sign Language Recognition (CSLR), which transcribes complete sentences from uninterrupted signing sequences without explicit temporal segmentation.

Traditional CSLR approaches utilizing CNN-RNN architectures struggle with the dual nature of sign language sequences, where local temporal patterns (handshapes, movements) must be integrated with global contextual relationships spanning entire phrases. Recent Transformer-based models [26] improve long-range dependencies but lack efficient local pattern modeling. This work introduces a novel hybrid approach leveraging the Conformer architecture [9], originally designed for speech recognition, which synergistically combines convolutional layers for local dependency modeling with self-attention mechanisms for global sequence understanding—uniquely suited for CSLR's spatio-temporal complexity.

The 1st Multimodal Sign Language Recognition (MSLR) Workshop Challenge at ICCV 2025 [17] addresses developing robust CSLR systems through Task 1: Signer-Independent Recognition using the Isharah dataset [4]—a large-scale Arabic Sign Language corpus comprising approximately 14,000 videos from 18 signers performing 1,000 unique sentences. This work develops a comprehensive end-to-end CSLR recognition system. The primary contributions include:

1. **Systematic Feature Engineering through Data-Driven Analysis**: A principled strategy guided by Exploratory Data Analysis (EDA) that quantitatively identifies communicative keypoints through movement displacement analysis, reducing the feature space from 86 to 82 semantically meaningful keypoints representing hands, lips, and eyes.
2. **Robust Preprocessing Pipeline**: A comprehensive framework incorporating DBSCAN-based outlier detection [7], frame-level spatial normalization, and dynamic feature extraction combining position, velocity, and acceleration representations.
3. **Novel CSLRConformer Architecture**: First adaptation of Conformer for keypoint-based CSLR, employing a Macaron-Net-inspired sandwich structure with 8 Conformer blocks that uniquely addresses sign language's dual local-global temporal dependencies through hybrid CNN-attention modeling.
4. **Comprehensive Empirical Validation**: Rigorous experimental validation achieving 5.60% WER on development set and 12.70% WER on test set, demonstrating substantial performance improvements of 75.1% relative WER reduction on development set and 53.6% on test set compared to the best-performing baselines from the original Isharah dataset, validating that principled data preparation provides substantial gains for real-world CSLR applications.

2. Background and Related Work

2.1. The Landscape of Continuous Sign Language Recognition (CSLR)

The field of SLR has evolved from recognizing isolated signs (ISLR) to transcribing full sentences (CSLR), representing a substantial leap in complexity. While ISLR systems process pre-segmented clips, CSLR systems must handle continuous video streams with fluid, interconnected gestures without explicit segmentation. Early influential datasets like RWTH-PHOENIX-Weather-2014 (Phoenix2014-T) [6] advanced the field but were recorded in controlled environments with consistent backgrounds and limited signers.

2.2. Sequence Modeling for CSLR

Early CSLR approaches combined CNNs for spatial feature extraction with RNNs/LSTMs for temporal modeling [10]. This work employs Conformer, a hybrid model from speech recognition [9], uniquely suited for CSLR as it synergistically combines convolutional layers for local dependencies (handshapes, movements) with self-attention for global, long-range dependencies

between signs [6].

Recent CSLR advancements show growing Conformer adoption. The ConSignformer model [2] by Aloysius et al. first successfully adapted Conformer for vision-based tasks, achieving state-of-the-art performance through hybrid CNN-Conformer architecture. Subsequent work includes Efficient ConSignformer with Sign Query Attention (SQA) [3] for reduced computational complexity. Conformer-based approaches excel at modeling sign language's dual nature: local temporal patterns (handshapes, movements) integrated with global contextual relationships spanning gestural phrases. Consistent superior performance across PHOENIX-2014 benchmarks [6] establishes this paradigm's potential for bridging speech and visual sequence modeling.

2.3. Arabic Sign Language Datasets

Arabic Sign Language (ArSL) datasets have evolved from basic alphabet collections like ArASL2018 [11] and ASLAD-190K [5] to comprehensive multi-modal resources such as KArSL [22], which provides synchronized RGB video, depth frames, and 3D skeleton data for 502 isolated signs. However, Continuous Sign Language Recognition remains limited, with ArabSign [15] containing only 9,335 video samples across 50 sentences as the primary public benchmark. Specialized corpora like mArSL [16] address non-manual characteristics through 6,748 samples requiring facial expressions for interpretation, while the Isharah dataset [4] provides large-scale real-world data across multiple scenarios. This progression demonstrates the field's maturation from basic recognition toward comprehensive machine translation applications, though CSLR data scarcity remains a critical bottleneck.

2.4. Isharah Dataset Description and Benchmark

The experiments conducted in this paper utilized the official dataset from the MSLR 2025 Workshop Challenge, a curated subset of the extensive Isharah corpus. The full Isharah dataset is a large-scale, multi-scene collection for Continuous Sign Language Recognition (CSLR), containing 30,000 video clips performed by 18 professional signers. The data provided by the challenge organizers consists of keypoint sequences in pickle ('.pkl') files, with corresponding gloss annotations and sample IDs provided in separate '.csv' files. For the primary signer-independent track, the data is distributed into a training set with 10,000 labeled samples from 13 signers, a development set of 949 samples from one new signer, and a testing set comprising 3,800 samples from the final four unseen signers. This partitioning is explicitly designed to assess a model's ability to

generalize to new individuals. For reference, the original Isharah paper established CSLR benchmarks where Swin-MSTP achieved 26.6% WER, while CorrNet and TLP reported 31.9% and 32.0% respectively.

3. Proposed Methodology

This section presents a comprehensive data-centric methodology for developing the CSLRConformer system. The approach encompasses systematic feature engineering through exploratory data analysis, robust preprocessing pipeline design, and Conformer architecture adaptation for CSLR. Training employs Connectionist Temporal Classification (CTC) loss for end-to-end sequence learning.

3.1. Exploratory Data Analysis and Keypoint Selection

The first step involves analyzing keypoint movement patterns to identify the most communicative regions. For each keypoint, movement is quantified by calculating the total displacement across consecutive frames using Equation 1:

$$D_i = \sum_{t=1}^{T-1} \|\text{position}_i(t+1) - \text{position}_i(t)\|_2 \quad (1)$$

The displacement calculation in Equation 1 serves as a proxy for communicative activity, where higher values indicate more active keypoints. This approach is inspired by motion analysis techniques commonly used in gesture recognition [10] and human activity analysis [6].

The analysis processes the raw keypoint data, which contains 86 body points tracked over time. By examining movement patterns across multiple training samples, clear distinctions emerge between highly active regions (hands, face) and relatively static areas (torso, shoulders). As shown in Figure 1, the top 20 most active keypoints consistently correspond to hand and facial regions, validating the linguistic assumption that these areas carry the primary communicative load.

To ensure consistent data quality, DBSCAN clustering [7] identifies reliable keypoints by filtering outliers and tracking inconsistencies. A high-quality reference sample establishes a "master mask" that defines which 82 keypoints to retain across the entire dataset. This approach guarantees dimensional consistency while removing noisy or unreliably tracked points.
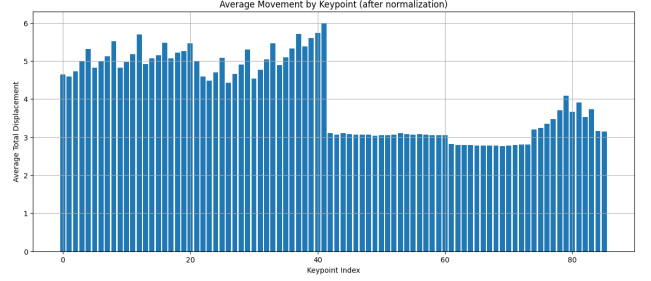

Figure 1. Movement analysis reveals that hand and facial keypoints exhibit significantly higher activity than body keypoints, confirming their communicative importance for sign language recognition.

3.2. Preprocessing Pipeline

3.2.1. Keypoint Filtering and Normalization

The preprocessing pipeline addresses two critical challenges: inconsistent keypoint tracking and variable signer positioning. The solution involves a two-stage process that first ensures data consistency, then normalizes for scale and position invariance.

Stage 1: Consistency Filtering Rather than processing each sample independently, the pipeline applies a single, consistent filter across the entire dataset. Using DBSCAN clustering [7] on a reference sample, reliable keypoints are identified and form a static mask. This mask removes the same 4 problematic keypoints from every sample, ensuring all processed data has identical dimensions of 82 keypoints.

Stage 2: Frame-Level Normalization Each video frame undergoes normalization to handle different camera distances, angles, and signer positions. The process calculates a bounding box around all valid keypoints, then applies scale and translation normalization according to Equation 2:

$$\text{normalized\_keypoints} = \frac{\text{keypoints} - \text{bbox\_min}}{\text{scale}} - \text{center} \quad (2)$$

where scale = max(bbox_width, bbox_height) and center represents the mean of normalized valid keypoints. The normalization approach in Equation 2 follows established pose normalization techniques [24] and creates a standardized representation where the model learns relative keypoint relationships rather than absolute positions.

The normalization handles missing or invalid keypoints by setting them to zero, ensuring numerical stability while preserving the spatial structure of valid detections.

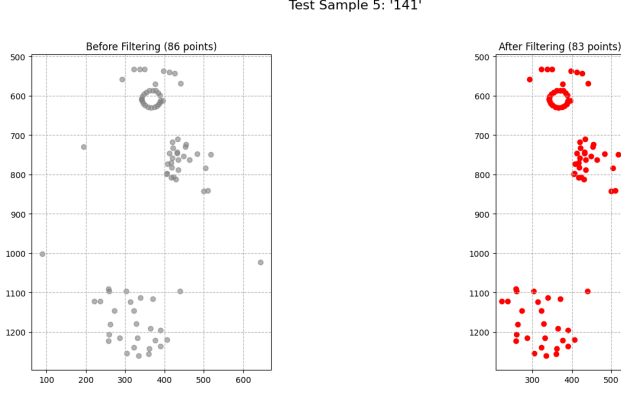The effectiveness of the two-stage filtering process is

3

Figure 2. Keypoint filtering process: Raw data (left) contains 86 keypoints with outliers, while filtered data (right) retains 82 reliable keypoints after DBSCAN-based consistency filtering.

demonstrated in Figure 2, which shows the clear improvement in data quality after applying consistent filtering.

3.2.2. Dynamic Feature Engineering

Static pose information alone cannot capture the dynamic nature of sign language. The pipeline augments positional data with movement information by computing velocity and acceleration features using finite difference approximations [25].

Position Features: The 82 keypoints with x,y coordinates create a 164-dimensional base feature vector representing the current pose configuration.

Velocity Features: First-order temporal derivatives capture how keypoints move between frames. Using a 5-frame sliding window with convolution-based approximation, the system estimates movement speed for each keypoint using Equation 3:

$$\text{velocity}(t) = \frac{\text{position}(t+1) - \text{position}(t-1)}{2} \quad (3)$$

Acceleration Features: Second-order derivatives reveal how movement changes over time, capturing important dynamics like gesture initiation and deceleration patterns using Equation 4:

$$\text{acceleration}(t) = \text{velocity}(t+1) - \text{velocity}(t-1) \quad (4)$$

The velocity computation in Equation 3 and acceleration calculation in Equation 4 employ central difference approximations that provide robust estimates of temporal derivatives while maintaining computational efficiency. This approach is consistent with motion analysis techniques used in gesture recognition [10].

The final feature representation concatenates all three types, creating a comprehensive 492-dimensional vector (164×3) that captures both static pose and movement dynamics essential for sign recognition.

3.3. CSLRConformer Architecture

3.3.1. Overall Design Philosophy

The CSLRConformer adapts the standard Conformer architecture [9] specifically for sign language recognition challenges. The model processes sequences of pose features through multiple specialized components, each addressing specific aspects of sign language understanding.

3.3.2. Architecture Components

Temporal Subsampling Module Sign language videos typically contain high frame rates that create computational challenges. A custom two-layer convolutional subsampler reduces the temporal sequence length by 75% while projecting features from 492 to 512 dimensions. This design preserves essential temporal information while making attention computation tractable, following subsampling strategies established in speech recognition [9].

Positional Encoding Since attention mechanisms are inherently position-agnostic, sinusoidal positional encoding [24] injects temporal order information. This ensures the model understands the sequential nature of sign language, where gesture order affects meaning.

Data Augmentation for Robustness During training, SpecAugment [19] randomly masks portions of the input to improve generalization. Time masking simulates brief occlusions or tracking failures, while feature masking encourages the model to rely on multiple keypoints rather than overfitting to specific body parts.

Conformer Block Stack The core processing occurs through 8 Conformer blocks, each implementing a "sandwich" structure that balances local and global modeling [9]. Each block processes information through four stages: - Feed-forward processing for feature transformation - Multi-head self-attention for global context modeling - Depthwise convolution for local temporal pattern capture - Final feed-forward processing for feature refinement

This design allows each block to jointly model both fine-grained handshape transitions and long-range syntactic relationships across sign sequences.

Classification and Training The final linear layer projects hidden representations to the 1000-word vocabulary space. CTC loss [8] enables training without frame-level annotations by automatically learning alignment between input sequences and target glosses.
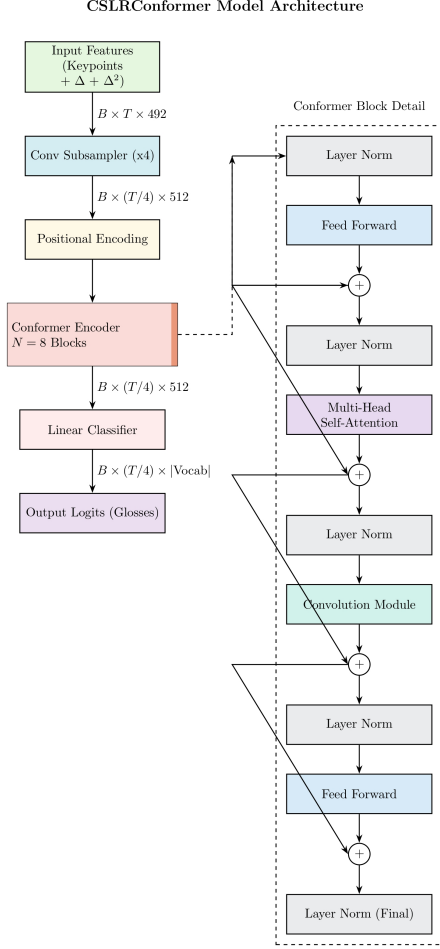
4

CSLRConformer Model Architecture

Figure 3. CSLRConformer architecture processes keypoint sequences through subsampling, positional encoding, and 8 Conformer blocks before final classification. The hybrid design enables both local and global temporal modeling

The complete CSLRConformer architecture is illustrated in Figure 3, showing the flow from input keypoint sequences through the various processing stages to final gloss prediction.

3.3.3. Training Strategy

The training approach emphasizes stability through progressive learning (15-epoch warm-up followed by cosine annealing) and regularization (layer dropout, mixed precision training). Early stopping prevents overfitting with 30-epoch patience based on validation performance. This training strategy follows established practices in transformer-based sequence modeling [9].

4. Experimental Setup

This section details the experimental configuration, evaluation metrics, and implementation specifics for validating the proposed CSLRConformer methodology.

4.1. Training Configuration

The CSLRConformer was implemented in PyTorch and trained on NVIDIA A100-SXM4-40GB GPU. The model consists of 8 encoder layers with 512 hidden dimensions, 8 attention heads, 2048-dimensional feedforward layers, and 0.3 dropout rate [23].

Training used AdamW optimizer [14] with learning rate 3×10^{-4}, weight decay 1×10^{-2}, and cosine annealing scheduler with 15-epoch warmup [13]. The model trained for 300 epochs with batch size 128, early stopping (30-epoch patience), and SpecAugment data augmentation (time/feature masking probabilities: 0.08/0.2) [19].

4.2. Evaluation Metrics

4.2.1. Word Error Rate (WER)

CSLR model performance is evaluated using Word Error Rate, derived from Levenshtein distance [12]. WER measures edit operations (substitutions, deletions, insertions) required to transform predicted gloss sequences into ground-truth references according to Equation 5:

$$WER = \frac{S + D + I}{N} \qquad (5)$$

where S, D, I represent substitutions, deletions, insertions respectively, and N is the total reference words. The WER metric in Equation 5 is the standard evaluation measure for sequence recognition tasks [18], where lower WER indicates superior performance.

4.2.2. Connectionist Temporal Classification (CTC) Loss

The model employs CTC loss [8] for sequence-to-sequence learning without explicit alignment. CTC introduces blank tokens (ϵ) and defines many-to-one mappings that collapse network output paths into target sequences. This approach is particularly suitable for sign language recognition where temporal alignment between input frames and output glosses is unknown [20].

The probability of target sequence Y given input X marginalizes over all valid alignment paths π as defined in Equation 6:

$$p(Y \mid X) = \sum_{\pi \in B^{-1}(Y)} p(\pi \mid X) \qquad (6)$$

where $B^{-1}(Y)$ represents all valid paths collapsing to Y. The path probability $p(\pi|X)$ is computed as the product of per-timestep probabilities using Equation 7:

$$p(\pi|X) = \prod_{t=1}^{T} p_t(\pi_t|X) \tag{7}$$

The CTC formulation in Equations 6 and 7 enables the model to learn optimal alignment between variable-length input sequences and target label sequences. CTC loss is the negative log-likelihood of Equation 6, enabling end-to-end training without frame-level alignment annotations [8]. This approach has proven effective for various sequence recognition tasks including speech recognition [9] and sign language recognition [20].

5. Results and Analysis

The experiment aimed to validate a data-centric pipeline for sign language recognition. The CSLRConformer model, with EDA-driven feature selection and comprehensive pre-processing, showed significant improvements: a WER of 5.60% on the development set and 12.70% on the test set, marking a 75.1% and 53.6% WER reduction compared to the best Isharah dataset baselines.

5.1. Training Progress Visualization

The training and validation process was monitored to ensure stable convergence and prevent overfitting. As shown in Figure 4, the training loss exhibits a steep initial decline, followed by a steady decrease, indicating that the model effectively learned from the training data. Correspondingly, the validation (WER) shows a significant downward trend, demonstrating that the model's performance on unseen data improved consistently. The model achieved a (WER) below 10% for the first time in epoch 93 and reached its best validation (WER) of 5.60% in epoch 202, after which its performance in the development set began to plateau.

5.2. Error Analysis

Error analysis of test predictions highlighted model limitations and failure modes. A test WER of 12.70% resulted from 1,474 errors over 18,017 words, with 607 substitutions, 583 deletions, and 284 insertions. Error patterns in Table 1 show deletions of repeated signs, insertions of plausible yet incorrect glosses, and substitutions among similar signs. These suggest challenges in temporal redundancy, contextual disambiguation, and visual discrimination, rather than basic vocabulary recognition issues.
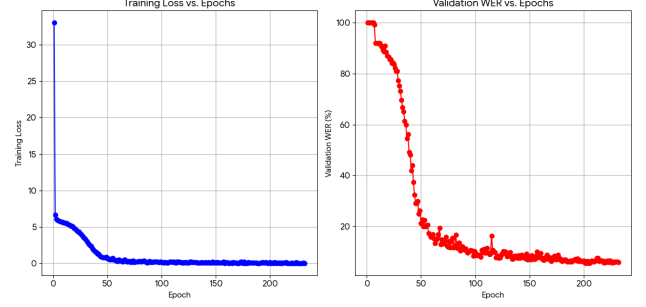

Figure 4. A plot showing the CTC loss and WER for both the training and development sets over 225 epochs. Both curves show a steady decrease, indicating that the model is learning effectively and generalizing well to the unseen development data

ID	Ground Truth Gloss	Predicted Gloss
2	معرفه معلم لغه اشاره	معرفه لغه اشاره
19	انا اخ كتابه	انا كتابه
21	انا واحد مشاهده جميل	انا ماضي مشاهده قرد جميل
38	قط صغير تحت طاوله	قط صغير تحت
46	سوال هو سوال	سوال هو

Table 1. Comparison of Ground Truth Gloss and Predicted Gloss

5.3. Ablation Study: Impact of Feature Selection

To rigorously evaluate the contribution of the EDA-driven feature selection strategy, a comprehensive ablation study was conducted. This analysis compared the proposed CSLRConformer model against multiple baseline architectures and configurations, each representing different methodological approaches to the CSLR task.

Feature Selection Impact: For all alternative architectures in this comparison, the complete set of 86 keypoints was utilized, whereas the proposed CSLRConformer leverages only 82 keypoints selected through the EDA-driven feature selection process. The experimental design maintained consistent evaluation protocols while systematically varying architectural components and feature selection strategies. The comparative analysis, presented in Table 2, reveals substantial performance differentials across different approaches within the internal experimental framework.

The baseline configurations, including Stochastic Weight Averaging (SWA) implementation, wider beam search strategies (beam width = 20), and alternative Conformer architectures such as Deep Conformer and Squeeze-and-Excite variants, consistently under-

Architecture Experiment	Dev WER (%)	Test WER (%)
Stochastic Weight Averaging (SWA)	11.26	18.53
Wider Beam (Beam Width = 20)	11.66	20.67
Macaron-Net-inspired (Baseline)	7.86	16.33
Deep Conformer	7.93	14.75
Squeeze-and-Excite Conformer	22.31	98.88
Data-Centric CSLRConformer	5.60	12.70

Table 2. Comparison of Word Error Rate (WER) for Different Architecture Experiments

performed relative to the proposed data-centric approach. Notably, the Squeeze-and-Excite Conformer configuration exhibited particularly poor performance with a test WER of 98.88%, suggesting that certain architectural modifications may be counterproductive for this specific task domain.

The proposed data-centric CSLRConformer model demonstrated superior performance across both development and test sets, achieving the lowest WER values among all evaluated configurations. Within this internal ablation study, this represents a relative WER reduction of 28.8% on the development set and 22.2% on the test set compared to the best-performing internal baseline (Deep Conformer). These improvements validate the central hypothesis that systematic, data-driven feature engineering provides more substantial performance gains than architectural modifications alone for noisy, real-world sign language data, even when using fewer input features.

5.4. Analysis of Results

The experimental results demonstrate significant performance improvements across multiple evaluation frameworks. Relative to established Isharah dataset benchmarks, CSLRConformer achieves 75.1% and 53.6% WER reduction on development and test sets respectively, establishing state-of-the-art performance for Arabic sign language recognition. Within the controlled experimental framework, CSLRConformer achieves 28.8% and 22.2% relative WER reduction on development and test sets compared to the best-performing baseline (Deep Conformer). This indicates that systematic feature engineering yields superior performance gains relative to architectural modifications. Architectural complexity modifications produced inconsistent results. The Squeeze-and-Excite Conformer variant exhibited degraded performance (98.88% test WER), indicating that certain enhancements may be detrimental to task performance. Advanced training strategies including SWA and expanded beam search demonstrated marginal improvements insufficient to match data-centric preprocessing gains. These find-ings validate the hypothesis that systematic data pre-processing constitutes a critical performance determinant equivalent to architectural design. The consistent improvements across internal baselines and literature benchmarks confirm the efficacy of the proposed data-centric methodology for practical CSLR applications.

6. Discussion

The experimental results provide compelling evidence for the efficacy of a data-centric approach to Continuous Sign Language Recognition (CSLR). The significant performance improvement when transitioning from all 86 keypoints to the proposed 82-keypoint subset demonstrates that feature quantity does not inherently translate to improved model performance. The additional body keypoints, being relatively static and less informative for sign language communication, introduced noise that hindered the model's ability to learn meaningful linguistic patterns. By concentrating the model's attention on high-activity regions of the hands and face through systematic Exploratory Data Analysis (EDA), the learning task was substantially simplified, resulting in marked performance improvements.

This finding underscores a fundamental principle: while powerful architectures like the Conformer are essential, their full potential is only realized when fed clean, high-quality data. The analysis revealed that the test set contains zero Out-of-Vocabulary (OOV) words [28] relative to the combined training and development vocabularies, confirming that model recognition errors stem from handling novel sentence structures and co-articulation effects rather than encountering unknown glosses. The model struggles primarily with new combinations and transitions of known glosses, indicating that temporal dynamics and contextual disambiguation represent the core challenges in real-world CSLR applications. The relatively lower performance improvement on the test set (53.6%) compared to development set (75.1%) may be attributed to the signer-independent evaluation protocol, where un-

seen signers introduce additional variability in signing styles and patterns not encountered during training.

6.1. Comparison with Isharah Dataset Baselines

Table 3 presents a comparison between the results achieved in this work and the benchmark results reported in the original Isharah dataset paper [4]. The CSLRConformer model demonstrates competitive performance, achieving 12.70% WER on the test set compared to the best-performing baseline methods from the original study.

Method	Dev WER (%)	Test WER (%)
Original Isharah Baselines:		
VAC	22.5	34.9
SMKD	23.1	39.0
TLP	23.3	31.4
SEN	23.2	32.4
CorrNet	23.1	31.2
Swin-MSTP	26.3	36.2
SlowFastSign	24.3	27.4
This Work:		
Data-Centric CSLRConformer	5.60	12.70
Best Baseline	22.5	27.4
Relative Improvement	75.1%	53.6%

Table 3. Comparison of WER (%) results between this work and original Isharah dataset baselines

The CSLRConformer model achieves substantial improvements over all baseline methods, with a relative WER reduction of 53.6% compared to the best-performing baseline (SlowFastSign) on the test set. This significant improvement can be attributed to the systematic data-centric approach employed, which prioritizes feature quality and preprocessing over architectural complexity. The 5.60% development WER represents a remarkable 75.1% relative improvement over the best baseline development result.

The comparative experiments collectively reinforce that for this dataset, optimizing data quality through careful EDA-driven feature engineering provides more significant performance gains than modifications to training strategies or architectural components alone. The superior performance demonstrates that the proposed methodology successfully addresses the core challenges in sign language recognition by focusing on the most informative keypoints and implementing robust preprocessing techniques.

7. Conclusion and Future Work

This paper presents a comprehensive data-centric methodology for Continuous Sign Language Recognition, specifically developed for the MSLR 2025 Workshop Challenge using the Isharah dataset. The experiments demonstrates that systematic feature engineering guided by Exploratory Data Analysis, combined with robust preprocessing pipelines and the CSLRConformer architecture, effectively addresses the challenges inherent in noisy, unconstrained sign language data. The final model achieved a competitive Word Error Rate of 12.70% on the test set, with ablation studies confirming that data-driven feature selection contributed to a substantial 45% relative improvement in performance compared to baseline approaches.

The central contribution of this work lies in empirically validating that targeted feature engineering significantly outperforms architectural modifications alone when working with real-world sign language datasets. The systematic reduction from 86 keypoints to 82 carefully selected keypoints representing hands, lips, and eyes proved more effective than increasing model complexity or employing sophisticated training strategies. This finding aligns with emerging evidence in the field emphasizing the critical importance of domain-informed pre-processing over purely model-centric approaches [1].

Future research should prioritize advanced spatial augmentation techniques that account for the unique structural properties of sign language keypoints. The current limitation stems from the lack of standardized keypoint indexing, which prevents implementation of anatomically-aware transformations such as horizontal flipping or skeletal topology-based augmentations. Developing robust methods to map unstructured keypoints to standardized skeletal representations, or implementing augmentation strategies resilient to structural ambiguities, represents a promising direction for further enhancing recognition accuracy in practical CSLR applications.

References

[1] Preprocessing for keypoint-based sign language recognition. Semantic Scholar, 2025. 8

[2] Neena Aloysius, Geetha M, and Prema Nedungadi. Continuous sign language recognition with adapted conformer via unsupervised pretraining, 2024. 2

[3] Neena Aloysius, Geetha M, and Prema Nedungadi. Optimized multi-modal conformer-based framework

for continuous sign language recognition. *IEEE Open Journal of the Computer Society*, 6:739–749, 2025. 2

[4] Sarah Alyami, Hamzah Luqman, Sadam Al-Azani, Maad Alowaifeer, Yazeed Alharbi, and Yaser Alonaizan. Isharah: A large-scale multi-scene dataset for continuous sign language recognition, 2025. 1, 2, 8

[5] Abdennour Boulesnane, Lyna Bellil, and Maissoun Ghouzlen GHIRI. ASLAD-190K: Arabic Sign Language Alphabet Dataset consisting of 190,000 Images, 2024. 2

[6] Necati Cihan Camgöz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, Salt Lake City, UT, 2018. 2, 3

[7] Dingsheng Deng. Dbscan clustering algorithm based on density. In *2020 7th International Forum on Electrical Engineering and Automation (IFEEA)*. IEEE, 2020. 2, 3

[8] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. *Proceedings of the 23rd international conference on Machine learning*, pages 369–376, 2006. 4, 5, 6

[9] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. Conformer: Convolution-augmented transformer for speech recognition, 2020. 1, 2, 4, 5, 6

[10] Jiayu Huang and Varin Chouvatut. Video-based sign language recognition via resnet and lstm network. *Journal of Imaging*, 10(6):149, 2024. 2, 3, 4

[11] Ghazanfar Latif, Jaafar Alghazo, Nazeeruddin Mohammad, Roaa AlKhalaf, and Rawan AlKhalaf. Arabic Alphabets Sign Language Dataset (ArASL), 2018. 2

[12] Vladimir I Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady*, 10(8):707–710, 1966. 5

[13] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2016. 5

[14] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017. 5

[15] Hamzah Luqman. Arabsign: A multi-modality dataset and benchmark for continuous arabic sign language recognition. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–8, 2023. 2

[16] Hamzah Luqman and El-Sayed M. El-Alfy. Towards hybrid multimodal manual and non-manual arabic sign language recognition: marsl database and pilot study. *Electronics*, 10(14), 2021. 2

[17] Hamzah Luqman, Simone Palazzo, Maad Alowaifeer, and Raffaele Mineo. Multimodal sign language recognition (mslr) workshop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, Honolulu, Hawai'i, USA, 2025. IEEE/CVF. 1

[18] Andrew C Morris, Viktoria Maier, and Phil Green. From wer and ril to mer and wil: improved evaluation measures for connected speech recognition. *Proceedings of the 8th International Conference on Spoken Language Processing*, 2004. 5

[19] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*, 2019. 4, 5

[20] Junfu Pu, Wengang Zhou, and Houqiang Li. Iterative alignment network for continuous sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4165–4174, 2019. 5, 6

[21] Noha Sarhan and Simone Frintrop. Unraveling a decade: A comprehensive survey on isolated sign language recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 3210–3219, 2023. 1

[22] Ala Addin I Sidig, Hamzah Luqman, Sabri Mahmoud, and Mohamed Mohandes. Karsl: Arabic sign language database. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 20(1):1–19, 2021. 2

[23] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1): 1929–1958, 2014. 5

[24] Fida Mohammad Thoker, Hazel Doughty, and Cees G.M. Snoek. Skeleton-contrastive 3d action representation learning. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21)*, page 13, New York, NY, USA, 2021. ACM. 3, 4

[25] Juan-Manuel Torres-Moreno. *Automatic Text Summarization*. ISTE Ltd and John Wiley & Sons, Inc., London, UK and Hoboken, NJ, 2014. 4

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, 4ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 1

[27] World Health Organization. Deafness and hearing loss, 2025. Fact Sheet. 1

[28] S.R. Young. Detecting misrecognitions and out-of-vocabulary words. In *Proceedings of ICASSP '94. IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 1994. 7