SvedbergOpen
DISSEMINATION OF KNOWLEDGE
International Journal of Artificial Intelligence and Machine Learning
Publisher's Home Page: https://www.svedbergopen.com/
Research Paper Open Access

Transformer-Based Deep Learning Model With Fusion-Based Reranking For Word-Level Sign Language Recognition

Maricel A. Esclamado

University of Science and Technology of Southern Philippines, Cagayan de Oro City, 9000, Philippines

Abstract
Word-level sign language recognition remains challenging because of similarities among signs and the limited availability of balanced training data. This study proposes a Transformer-based deep learning model with fusion-based reranking for word-level sign language recognition using MediaPipe hand landmark sequences. A subset of the WLASL dataset was used in the analysis. Hand landmarks were extracted, preprocessed, and augmented before being used to train a Transformer encoder. To refine the predicted gloss rankings, clustering confidence scores obtained from K-means were combined with the Transformer's prediction probabilities through a fusion-based reranking strategy. Experimental results showed that reranking significantly improved Top-5, Top-10, and Top-20 accuracies, while Top-1 accuracy remained unchanged. These findings demonstrate that incorporating clustering information can effectively improve candidate ranking and enhance the overall performance of Transformer-based word-level sign language recognition.

Keywords: Transformer-based, Clustering, Fusion-based Reranking, Sign Language Recognition.

I. Introduction

Sign language recognition has become an important area of research because it offers a means of improving communication between deaf or hard-of-hearing individuals and those who do not understand sign language. Recent advances in computer vision and deep learning have enabled the development of vision-based recognition systems that can learn complex hand movements directly from video data without the need for wearable sensors (Ramya et al., 2024; Liang et al., 2023). Among these approaches, Transformer-based models have shown considerable promise due to their ability to capture long-range spatial and temporal relationships in sequential data.

Despite these advances, word-level sign language recognition remains a challenging task because of the high similarity among many signs, variations in signing styles, and the imbalance commonly found in available datasets. These challenges often lead to incorrect ranking of candidate predictions, particularly when multiple glosses exhibit similar hand configurations and movement patterns. Improving the ranking of candidate predictions therefore remains an important research problem.

This study proposes a Transformer-based deep learning model with fusion-based reranking for word-level sign language recognition. Hand landmark sequences extracted using MediaPipe are used as the model input to reduce computational complexity while preserving essential gesture information. In addition, clustering confidence scores are combined with the Transformer's prediction probabilities through a fusion-based reranking strategy to refine the ranking of candidate glosses. The proposed approach is evaluated using a subset of the WLASL dataset to determine its effectiveness in improving recognition performance. This study would like to answer the following research questions:

1. How effectively can a Transformer-based deep learning model recognize word-level sign language using MediaPipe hand landmark sequences?
2. Does the proposed fusion-based reranking strategy improve the Top-k recognition performance of the Transformer-based model compared with using the Transformer classifier alone?

II. Literature Review

Communication remains a significant challenge for deaf and hard-of-hearing individuals, particularly when interacting with people who are unfamiliar with sign language. To bridge this communication gap, researchers have developed Sign Language Recognition (SLR) systems that leverage computer vision and artificial intelligence to automatically interpret sign language gestures. Over the years, these systems have evolved from hardware-dependent solutions to more accessible vision-based approaches with machine learning and deep learning (ZainEldin et al., 2024).

Different visual representations have been explored to improve recognition performance. Initial studies commonly used raw RGB images or video frames as model inputs, extracting features through Convolutional Neural Networks (CNNs) before modeling temporal information with deep learning models such as Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU) networks. More recently, Transformer architectures have gained attention because of their ability to model long-range temporal dependencies through self-attention mechanisms, allowing them to process entire sign sequences more effectively than recurrent models. While image-based approaches have achieved promising results, they often require large computational resources and may be sensitive to background variations, lightning conditions, and other visual information. Mediapipe has become one of the most widely used frameworks for this purpose because it provides accurate real-time detection of hand, face, and pose landmarks while significantly reducing computational complexity. Rather than learning directly from raw images, models trained on landmark coordinates focus on the geometric structure and movement of the signer, making them more robust and efficient.

Several studies have demonstrated the effectiveness of Mediapipe landmarks for sign language recognition. Luna-Jiménez et al. (2023) demonstrated the potential of this approach by combining MediaPipe landmarks with Transformer-based architectures and reported improved recognition performance while enhancing model interpretability by identifying the body movements that can contributed to each prediction. Similarly, Ahmed et al. (2023) used distances between Mediapipe hand landmarks as features for machine learning classifiers and achieved high recognition accuracy, with the support vector machine outperforming other algorithms. These findings suggest that landmark-based representations preserve essential gesture information while reducing the computational burden associated with processing full images.

Building on these developments, the present study (Berepiki et al., 2025) adopts a landmark-based approach by utilizing MediaPipe to extract hand landmarks from sign language videos. The extracted landmark sequences are then modeled using a Transformer network to learn the temporal relationships among gestures. By focusing on landmark coordinates rather than raw images, the proposed approach aims to improve computational efficiency while preserving the spatial and temporal characteristics necessary for accurate sign language recognition. This combination of MediaPipe landmarks and Transformer learning reflects the current direction of state-of-the-art SLR research and provides a practical framework for recognizing a large vocabulary of signs from the WLASL dataset.

Despite the considerable progress in SLR, challenges remain. Recognition performance can still be affected by differences in signing styles, occlusions, rapid hand movements, and the limited availability of large annotated datasets, particularly for regional sign languages. ZainEldin et al. (2024) emphasized the need for more robust recognition systems capable of generalizing across diverse signing conditions while maintaining computational efficiency. These challenges continue to motivate the development of more effective landmark-based deep learning models for real-world sign language recognition applications.

III. Methods

In this study, a hybrid sign language recognition framework that combines Transformer-based modeling with unsupervised clustering to improve gloss classification.

The data used in this analysis is a subset of the WLASL dataset which is the largest video dataset for word-level American Sign Language (ASL), containing 2,000 commonly used words in ASL (Li et al., 2020). Since some videos in the WLASL dataset are no longer available due to expired URLs, only the top 300 glosses with the highest number of available samples were included in the analysis. The split information provided in the WLASL json file was used to assign each video sample to the training, validation, or test set. To ensure that every gloss was represented across all three subsets, glosses that were not present in the training, validation, and test sets were excluded from the analysis.

The pipeline begins by extracting and preprocessing hand landmark sequences from sign language videos in the WLASL dataset before training a Transformer encoder to learn temporal representations of sign gestures through supervised learning. Each sign language video is first segmented into a sequence of frames. For every frame, hand landmarks are extracted using Mediapipe library in Python. Mediapipe can be used to extract 21 keypoints or landmarks from the hand (Gil-Martín et. al, 2025). The extracted landmarks for each frame are then stored in a sequence, where each sequence corresponds to a single sign

video. Since both left and right hands are considered, a total of 42 features are extracted per frame, representing the full set of hand landmarks. These features serve as the input representation for the Transformer encoder. To standardize the input data, videos longer than 132 frames were truncated to the fixed frame length, while videos with fewer frames than the minimum threshold of 62 frames were excluded from the dataset. The remaining shorter videos were padded with zeros to ensure that all input sequences had a uniform length. The frame length of 132 corresponds to the 95% percentile of the number frames of all the sign videos in the dataset.

Data Augmentation and Class Imbalance Mitigation

The dataset contains approximately 2,828 video samples representing 300 unique glosses. However, the dataset is imbalanced, with some glosses containing more video samples than others.

The videos were also collected from different signers and recorded under varying recording conditions, resulting in differences in hand position, movement amplitude, and signing style. To improve the model's ability to generalize to these variations, data augmentation addresses these challenges by generating additional training samples that preserve the semantic meaning of each sign while introducing realistic variations. Data augmentation was applied to the extracted hand landmark sequences using noise, shift, scaling, and speed transforms. These transformations were combined with model weight adjustments to address class imbalance in the dataset. Class weights, inversely proportional to the frequency of each gloss, were incorporated into the categorical cross-entropy loss during training. This encourages the model to pay greater attention to underrepresented glosses and helps reduce the tendency to favor classes with a larger number of training samples.

Transformer-based Encoder

A Transformer-based encoder was employed to learn the temporal dependencies present in sequential hand landmark data. Multiple iterations were done to find the appropriate values of hyperparameters considering the improvement in terms of loss and model performance. Early stopping is triggered if the validation loss showed no improvement for 10 consecutive epochs. Overfitting was monitored, and dropout rate and number of epochs were adjusted. The encoder consisted of two encoder layers with four multi-head self-attention heads and accepted fixed-length input sequences of 132 frames, each represented by 42 hand landmark features. The encoder output was connected to a classification head comprising two fully connected layers with 128 neurons and ReLU activation, followed by dropout (0.2) and L2 regularization (1e-4) to reduce overfitting before the softmax output layer for gloss classification. Performance was evaluated using Top-1, Top-5, Top-10, and Top-20 accuracy.

PCA and Clustering

The trained encoder performs gloss classification and also generates embedding representations for each input sequence. Principal Component Analysis (PCA) was applied to reduce feature dimensionality of the embedding vector while preserving most of the important information. PCA is a dimensionality reduction technique that transforms high-dimension data into lower-dimension while maintaining significant patterns and trends as possible (Jia et al., 2022). Using PCA, the dimension was reduced to 6 principal components, of which the cumulative explained variance exceeded a threshold of 95%. K-means clustering algorithm was then applied to the PCA-transformed embeddings to group samples with similar feature representations.

Fusion and Reranking

After the Transformer-based gloss classification and clustering, a fusion-based reranking strategy was applied to improve the ranking of the predicted glosses by combining the confidence scores from both approaches. For each test sample, the Transformer classifier produced a probability distribution over all glosses, from which the top-50 gloss candidates with the highest prediction probabilities were retained. Limiting the candidate set to the top 50 provided a balance between maintaining a sufficient number of possible glosses for reranking and keeping the computation efficient.

To incorporate clustering information, the Euclidean distances between each test embedding and the cluster centroids were converted into confidence scores using softmax function. A gloss-to-cluster mapping was then created from the training data by assigning each gloss to the cluster in which it most frequently occurred. During inference, this mapping allowed each Transformer candidate gloss to be linked to its corresponding cluster, enabling both the Transformer prediction probability scores and the cluster confidence to be considered during the reranking process.

For each of the top-50 Transformer predictions, the assigned gloss cluster was compared with the predicted clusters of the test sample. The corresponding cluster confidence score was assigned. A fusion score was then computed as a weighted linear combination of the Transformer prediction probability and the cluster confidence score:

$$F(g) = \alpha\, T(g) + \beta\, C(g)$$

where F(g) denotes the fusion score of gloss g, T(g) is the Transformer prediction probability, C(g) is the cluster confidence score, and α and β are weight coefficients**.** The candidate glosses were subsequently sorted in descending order according to their fusions scores to produced reranked prediction list. Performance was evaluated by comparing the reranked predictions against the ground-truth gloss labels using Top-1, Top-5, Top-10, and Top-20 accuracy.

IV. Results and Discussion

This section presents the results obtained from each stage of the proposed framework, beginning with hand landmark extraction. MediaPipe was used to identify and track the hand keypoints in every sign video, transforming the visual information into landmark sequences that served as input to the Transformer model. Sample outputs of the extracted hand landmarks are presented in Figure 1.

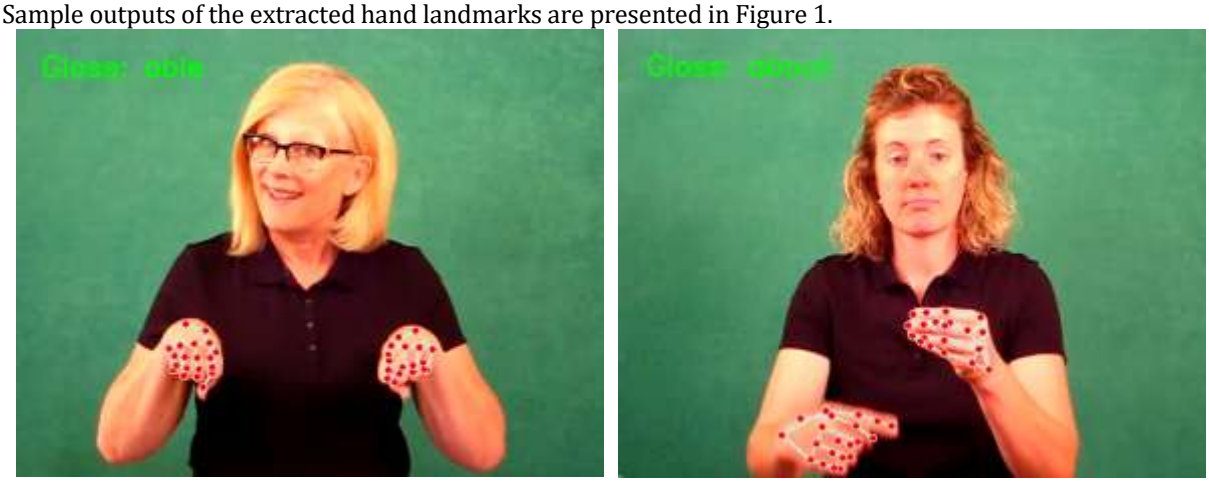

FIGURE 1. Hand Landmarks of sample videos

Both training and validation loss showed a downward trend as seen in Figure 2, with the greatest reduction occurring in the first five epochs. The losses continued to decrease gradually until around epoch 20, after which signs of overfitting began to emerge. Training and validation accuracies showed an overall upward trend as seen in Figure 3, indicating the model progressively learned more discriminative features for gloss classification. While the accuracy fluctuated slightly during training, its overall improvement suggests that the model maintained good generalization performance.

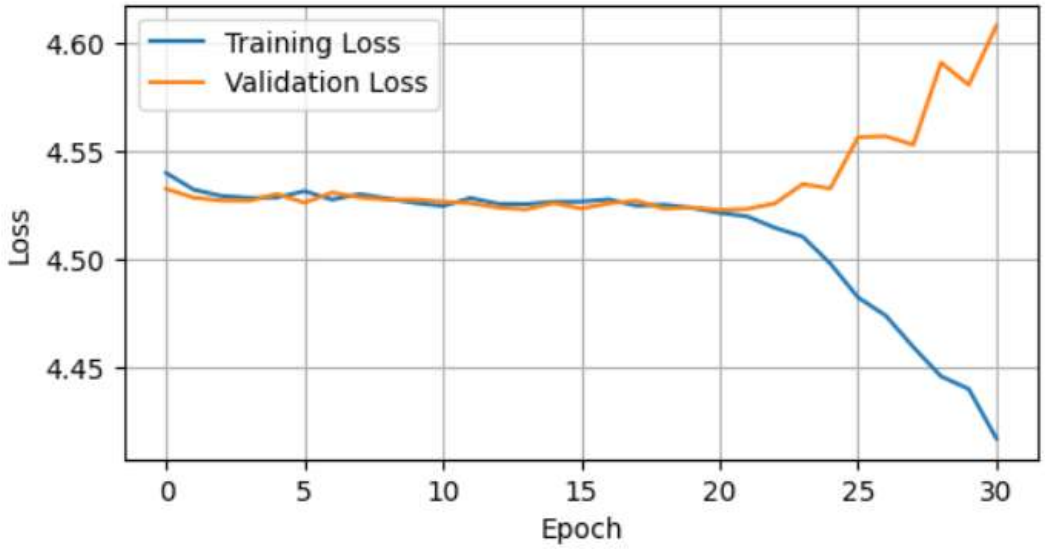

FIGURE 2. Training vs. Validation Loss

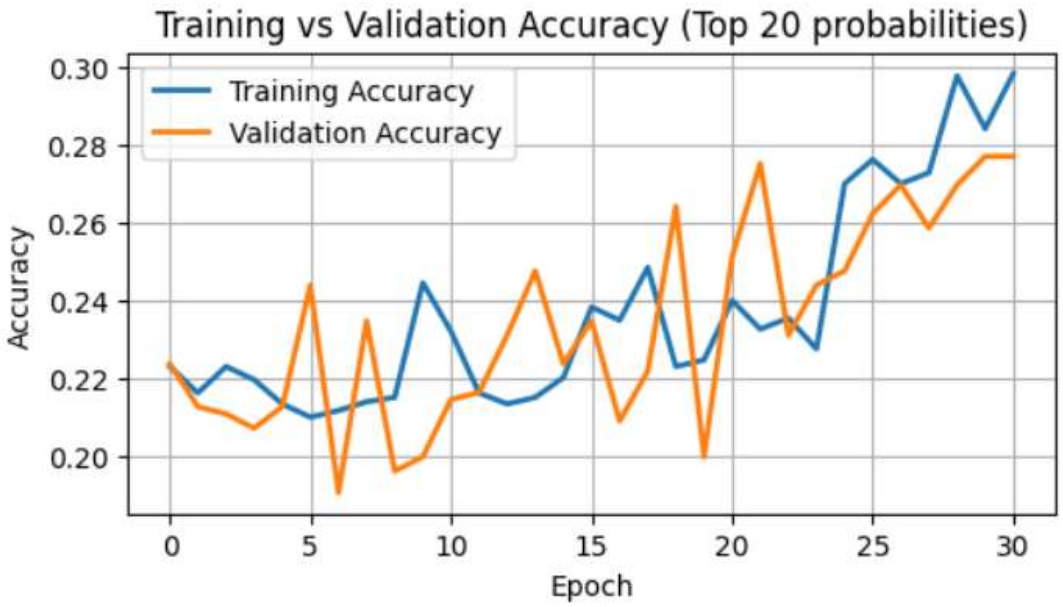

FIGURE 3. Training vs Validation Top-20 accuracy

Figure 4 presents the clustering results (k=7) on the PCA-reduced embeddings generated by the encoder. The clustering achieved a silhouette score of 0.56, indicating a moderate level of separation among the clusters. The clustering confidence scores were incorporated into the fusion-based reranking strategy. These scores provided additional information about the similarity of the predicted glosses in the embedding space and were combined with the Transformer's prediction probabilities during reranking.

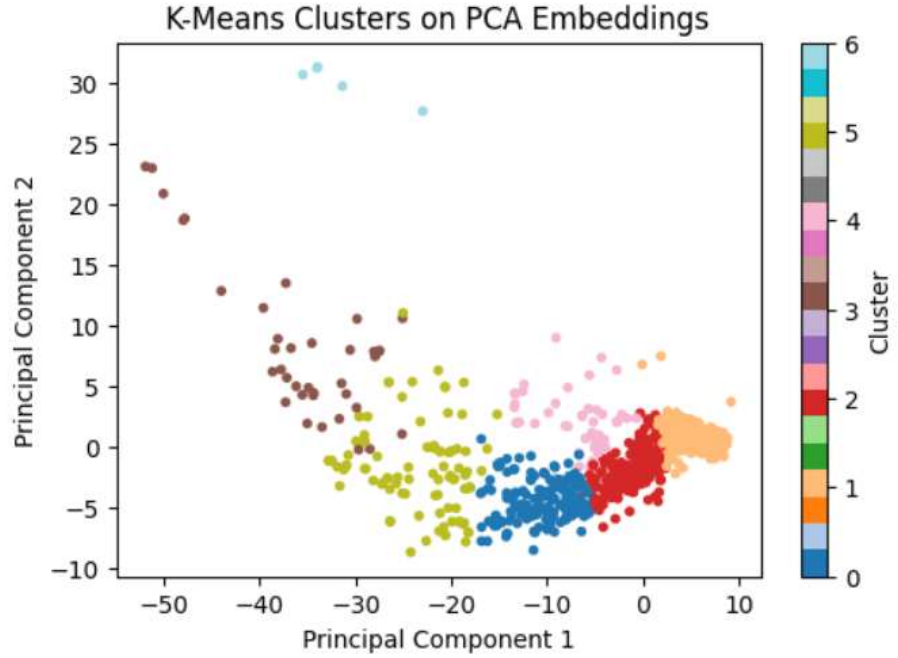

FIGURE 4. Clusters on PCA Embeddings

Table 1: Comparison of Transformer and Fusion-based Reranking Performance

Top-k accuracy	Transformer	Reranked	% Improvement
Top-1 accuracy	0.92%	0.92%	0%
Top-5 accuracy	0.92%	4.6%	400%
Top-10 accuracy	4.6%	13.2%	186.96%
Top-20 accuracy	13.2%	24.95%	89.02%

Table 1 compares the performance of the standalone transformer classifier with the fusion-based reranking approach. The results show that reranking had no effect on Top-1 accuracy, indicating that the highest-

ranked prediction generated by the Transformer was generally unchanged after the reranking process. This suggests that the cluster-based confidence was not sufficient to alter the first-ranked prediction. However, substantial improvements were observed for the higher-rank accuracy metrics. Top-5 accuracy increased from 0.92% to 4.6%, representing a 400% increase. Similarly, Top-10 accuracy improved from 4.6% to 13.2%, corresponding to an increase of 186.96%, while Top-20 accuracy increased from 13.2% to 24.95%, an improvement of 89.02%.

These findings indicate that the reranking strategy was effective in promoting the correct gloss to a higher position within the ranked candidate list, even when it was not initially the top prediction of the Transformer. By combining the prediction confidence from the Transformer with the cluster confidence obtained from the clustering, the reranking process refined the ordering of candidate glosses and improved retrieval performance. This enhancement is particularly beneficial in sign language recognition applications where multiple candidate predictions can be presented to users or used in subsequent language modeling and post-processing steps.

V. Conclusion

This study presented a Transformer-based deep learning model with fusion-based reranking for word-level sign language recognition using MediaPipe hand landmark sequences. The results demonstrated that the Transformer model was able to learn meaningful representations of sign language gestures from hand landmarks, addressing the first research question. However, the overall recognition accuracy remained relatively low because of the challenging nature of the task, which involved classifying a large number of glosses with limited training samples for many classes. In addition, some videos in the original WLASL dataset were no longer accessible due to expired URLs, reducing the amount of available training data and likely affecting the model's ability to learn robust representations.

The second research question examined whether the proposed fusion-based reranking strategy could improve recognition performance. Although reranking did not improve Top-1 accuracy, it substantially increased the Top-5, Top-10, and Top-20 accuracies by refining the ranking of candidate glosses using clustering confidence scores together with the Transformer's prediction probabilities. These findings suggest that clustering information can serve as useful complementary evidence for improving candidate ranking, particularly when the correct gloss is already among the Transformer's highest-ranked predictions.

While the proposed approach demonstrates the potential of combining Transformer-based learning with fusion-based reranking, there remains considerable room for improvement. Future work may explore larger and more complete datasets, additional landmark features such as body pose and facial landmarks, alternative Transformer architectures, and more advanced reranking or metric-learning techniques to further improve word-level sign language recognition performance.

References

1. Ahmed, S., Kar, S. K., & Basak, S. (2023, November). A novel approach for recognizing real-time american sign language (ASL) using the hand landmark distance and machine learning algorithms. In 2023 IEEE 9th International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE) (pp. 171-176). IEEE.
2. Berepiki, E. O., Ciunkiewicz, P., & Yanushkevich, S. (2025). Can a Lightweight Transformer Deliver a Robust Multimodal Sign Language Word Recognition?. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 4979-4986).
3. Gil-Martín, M., Marini, M. R., Martín-Fernández, I., Romero, S. E., & Cinque, L. (2025, February). Hand Gesture Recognition Using MediaPipe Landmarks and Deep Learning Networks. In ICAART (3) (pp. 24-30).
4. Jia, W., Sun, M., Lian, J., & Hou, S. (2022). Feature dimensionality reduction: a review. Complex & Intelligent Systems, 8(3), 2663-2693.
5. Li, D., Rodriguez, C., Yu, X., & Li, H. (2020). Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In Proceedings of the IEEE/CVF winter conference on applications of computer vision (pp. 1459-1469).
6. Liang, Z., Li, H., & Chai, J. (2023). Sign Language Translation: A Survey of Approaches and Techniques. Electronics, 12(12), 2678.
7. Luna-Jiménez, C., Gil-Martín, M., Kleinlein, R., San-Segundo, R., & Fernández-Martínez, F. (2023, October). Interpreting sign language recognition using transformers and MediaPipe landmarks. In Proceedings of the 25th International Conference on Multimodal Interaction (pp. 373-377).

8. Ramya, N., Kumar, G. K., Muskan, M. S., Chowdary, C. R., & Krishna, M. R. (2024). Enhancing Communication Accessibility: A Hand Gesture Recognition System for Deaf and Mute Individuals. Journal of Nonlinear Analysis and Optimization, 15(1).
9. ZainEldin, H., Gamel, S. A., Talaat, F. M., Aljohani, M., Baghdadi, N. A., Malki, A., & Elhosseini, M. A. (2024). Silent no more: a comprehensive review of artificial intelligence, deep learning, and machine learning in facilitating deaf and mute communication. Artificial Intelligence Review, 57(7), 188.