Real-Time Lightweight Sign Language Recognition on Hybrid Deep CNN-BiLSTM Neural Network with Attention Mechanism

Gulnur Kazbekova[1], Zhuldyz Ismagulova[2], Gulmira Ibrayeva[3],
Almagul Sundetova[4], Yntymak Abdrazakh[5], Boranbek Baimurzayev[6]
Khoja Akhmet Yassawi International Kazakh-Turkish University, Turkistan, Kazakhstan[1, 5, 6]
ALT University, Almaty, Kazakhstan[2]
Military Institute of the Air Defense Forces Named After Twice
Hero of the Soviet Union T.Ya. Bigeldinov, Aktobe, Kazakhstan[3]
Baishev University, Aktobe, Kazakhstan[4]

***Abstract*—Sign language recognition (SLR) plays a crucial role in bridging communication gaps for individuals with hearing and speech impairments. This study proposes a hybrid deep CNN-BiLSTM neural network with an attention mechanism for real-time and lightweight sign language recognition. The CNN module extracts spatial features from individual gesture frames, while the BiLSTM module captures temporal dependencies, enhancing classification accuracy. The attention mechanism further refines feature selection by focusing on the most relevant time steps in a sign sequence. The proposed model was evaluated on the Sign Language MNIST dataset, achieving state-of-the-art performance with high accuracy, precision, recall, and F1-score. Experimental results indicate that the model converges rapidly, maintains low misclassification rates, and effectively distinguishes between visually similar signs. Confusion matrix analysis and feature map visualizations provide deeper insights into the hierarchical feature extraction process. The results demonstrate that integrating spatial, temporal, and attention-based learning significantly improves recognition performance while maintaining computational efficiency. Despite its effectiveness, challenges such as misclassification in ambiguous gestures and real-time computational constraints remain, suggesting future improvements in multi-modal fusion, transformer-based architectures, and lightweight model optimizations. The proposed approach offers a scalable and efficient solution for real-time sign language recognition, contributing to the development of assistive technologies for individuals with communication disabilities.**

Keywords*—Sign language recognition; CNN-BiLSTM; attention mechanism; deep learning; gesture classification; real-time processing; assistive technology

I. Introduction

Sign language serves as a primary mode of communication for individuals with hearing and speech impairments, enabling them to interact effectively within society. However, barriers still exist due to the lack of widespread understanding and adoption of sign language by the general public. In this context, sign language recognition (SLR) plays a crucial role in bridging the communication gap between individuals with hearing disabilities and those who rely on spoken language [1]. The recent advancements in deep learning have paved the way for robust and efficient SLR systems, enhancing real-time communication through gesture-based interaction [2].

Traditional approaches to SLR have relied heavily on handcrafted feature extraction techniques, such as histogram of oriented gradients (HOG), scale-invariant feature transform (SIFT), and local binary patterns (LBP). While these methods have shown promise in controlled environments, their performance is often hindered by variations in lighting, occlusions, and user-specific differences in sign execution [3]. The emergence of deep learning techniques, particularly convolutional neural networks (CNNs), has revolutionized the field by enabling automatic feature extraction and classification with remarkable accuracy [4].

Recent research has demonstrated the effectiveness of CNN-based architectures for visual gesture recognition tasks, including sign language translation. However, CNNs alone lack the ability to capture temporal dependencies in sequential gesture data, which is essential for accurate recognition of continuous sign language sequences [5]. To address this limitation, hybrid deep learning models combining CNNs with recurrent neural networks (RNNs) or bidirectional long short-term memory (BiLSTM) networks have been proposed, allowing the extraction of both spatial and temporal features from sign language gestures [6]. The CNN component focuses on spatial feature extraction, while the BiLSTM module captures temporal dependencies in both forward and backward directions, thereby improving recognition accuracy [7].

Despite the promising results achieved through CNN-BiLSTM models, challenges remain in real-time SLR applications due to the computational complexity of deep learning networks. High processing requirements hinder their deployment on resource-constrained devices, such as mobile phones and embedded systems, which are essential for practical, real-world applications [8]. As a solution, lightweight neural network architectures have been explored, incorporating model compression techniques such as depthwise separable convolutions, pruning, and quantization to reduce computational overhead while maintaining high classification accuracy [9].

In addition to model efficiency, attention mechanisms have emerged as a powerful tool for enhancing performance in sequential data processing. The attention mechanism allows the model to selectively focus on relevant features within a sequence, improving temporal coherence in gesture recognition tasks [10]. When integrated into CNN-BiLSTM architectures, attention mechanisms enhance feature selection by emphasizing the most informative frames, thereby mitigating the impact of redundant or irrelevant information [11].

This study proposes a real-time, lightweight SLR system based on a hybrid deep CNN-BiLSTM architecture enhanced with an attention mechanism. The proposed framework is designed to achieve high recognition accuracy while minimizing computational costs, making it suitable for deployment on edge devices and mobile platforms [12]. The model leverages CNNs for extracting spatial features, BiLSTM networks for capturing bidirectional temporal dependencies, and an attention mechanism for focusing on salient information within sign sequences. By optimizing both accuracy and efficiency, this approach addresses the practical limitations of existing SLR systems [13], [14].

II. Related Works

A. Sign Language Recognition

Sign Language Recognition (SLR) has gained significant attention in recent years due to the increasing demand for assistive technologies aimed at bridging communication gaps between individuals with hearing disabilities and the wider community [15]. Various methods have been explored to achieve effective SLR, ranging from rule-based approaches to deep learning models [16]. Early approaches relied on handcrafted features extracted from gesture sequences, while modern techniques emphasize end-to-end learning using neural networks [17].

B. Traditional Methods

Before the advent of deep learning, traditional methods for SLR primarily relied on handcrafted feature extraction techniques such as HOG, SIFT, and LBP [18]. These methods extracted low-level features from hand gestures and used classification techniques such as Support Vector Machines (SVMs) and Hidden Markov Models (HMMs) to recognize signs [19]. While these approaches provided reasonable accuracy in controlled environments, they struggled with real-world variations such as occlusions, background noise, and different sign execution speeds [20].

C. Machine Learning Approaches

With the rise of machine learning, researchers began to explore data-driven approaches for SLR. Machine learning models, such as Random Forests and SVMs, demonstrated improved accuracy compared to traditional rule-based methods [21]. The introduction of artificial neural networks (ANNs) further enhanced recognition capabilities, allowing for automatic feature extraction and improved generalization to unseen sign variations [22]. However, these methods were still limited by their inability to effectively capture both spatial and temporal dependencies in sign language sequences [23].

D. Deep Learning for Sign Language Recognition

Deep learning has revolutionized SLR by providing powerful feature extraction and classification capabilities. Convolutional Neural Networks (CNNs) have been widely used for spatial feature extraction, achieving state-of-the-art performance in static sign recognition [24]. However, recognizing continuous sign language requires capturing temporal dependencies, which led to the integration of Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks into SLR frameworks [25]. More recently, BiLSTM networks have been employed to improve sequence modeling by considering both forward and backward temporal dependencies, leading to enhanced recognition accuracy [26]. Additionally, attention mechanisms have been incorporated into CNN-BiLSTM architectures to enhance feature selection and improve classification performance [27].

E. Challenges in Sign Language Recognition

Despite significant progress, several challenges remain in developing real-time and robust SLR systems. One major challenge is the variability in sign execution, including differences in speed, hand position, and occlusions [28]. Another challenge is the high computational cost of deep learning models, making it difficult to deploy them on edge devices and mobile platforms [29]. Addressing these challenges requires optimizing model architectures for efficiency while maintaining high recognition accuracy.

F. Research Gaps

While deep learning-based SLR systems have achieved remarkable success, there are still research gaps that need to be addressed. Existing models often require large labeled datasets, which are expensive and time-consuming to create [30]. Additionally, real-time processing remains a challenge due to the complexity of CNN-BiLSTM architectures [31]. Further research is needed to develop lightweight models that can operate efficiently on low-power devices without compromising recognition performance [32]. Moreover, integrating multi-modal inputs, such as depth and motion data, could enhance recognition robustness in real-world scenarios [33].

By addressing these gaps, future sign language recognition systems can be made more efficient, accurate, and accessible, ultimately improving communication for individuals with hearing impairments.

III. Problem Statement

The fundamental challenge in Sign Language Recognition (SLR) is achieving high-accuracy, real-time classification of gestures while maintaining computational efficiency. Given an input sequence of image frames $X = \{x_1, x_2, ..., x_T\}$, the goal is to predict the corresponding sequence of sign language labels $Y = \{y_1, y_2, ..., y_T\}$ such that:

$$P(Y \mid X) = \arg\max P(y_t \mid x_t, \theta) \quad (1)$$

where, θ represents the learned parameters of the model.

Traditional deep learning approaches rely on CNNs for spatial feature extraction and LSTMs for temporal dependencies. However, existing methods struggle with balancing recognition accuracy and real-time efficiency, especially in low-resource environments. Thus, a hybrid CNN-BiLSTM model with an attention mechanism is needed to enhance spatial-temporal feature extraction while maintaining lightweight computational costs.

IV. Materials and Methods

Developing an accurate and efficient Sign Language Recognition (SLR) system requires a well-structured methodology that encompasses dataset selection, preprocessing, model architecture, and training strategies. This section provides a comprehensive overview of the materials and methods employed in this study. First, the dataset used for training and evaluation is described, including its structure, distribution, and preprocessing techniques. Next, the proposed hybrid CNN- BiLSTM model with an attention mechanism is introduced, detailing its ability to extract spatial and temporal features from sign language gestures. The section further elaborates on the training process, including the optimization strategies, loss functions, and performance evaluation metrics utilized. Finally, implementation details, including computational resources and hyperparameter settings, are presented to ensure the reproducibility of the study.

A. Dataset

The Sign Language MNIST dataset is a widely used benchmark for static sign language recognition. It was designed as an adaptation of the MNIST dataset to facilitate research in sign language gesture classification [34]. The dataset consists of 27,455 grayscale images, each of size 28×28 pixels, representing 24 different hand gestures corresponding to the American Sign Language (ASL) alphabet. The dataset excludes the letters J and Z since these signs involve dynamic motion that cannot be effectively captured in static images.

The dataset is divided into two subsets: a training set of 27,455 images and a test set of 7,172 images, ensuring a structured approach to evaluating model performance. Each image represents a single hand gesture and is labeled with one of the 24 classes. The data is well-balanced across the different sign categories, enabling efficient training of deep learning models.

The simplicity of the dataset, coupled with its structured grayscale format, makes it an ideal benchmark for evaluating convolutional neural networks (CNNs) and hybrid deep learning architectures for sign language recognition. Fig. 1 provides a visual representation of sample images from the dataset, illustrating the variation in hand gestures and their corresponding labels.

Fig. 2 presents a visualization of the class distribution within the Sign Language MNIST dataset. The dataset comprises 24 distinct hand gesture classes, each representing a different letter in the American Sign Language (ASL) alphabet, excluding J and Z, which require motion. The histogram illustrates the number of samples per class, providing insight into the dataset's balance.

Fig. 1. Sample images of the applied dataset.

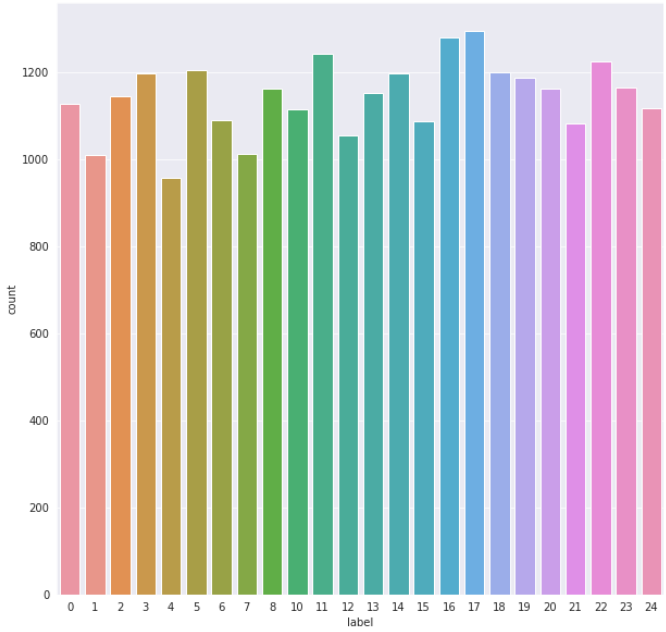

Fig. 2. Distribution of classes.

From Fig. 2, it can be observed that the dataset is relatively balanced, with each class containing approximately 1,000 to 1,250 samples. This balanced distribution is crucial for training deep learning models, as it minimizes the risk of class bias and ensures that all gestures receive equal representation during training. A well-distributed dataset allows for better generalization, reducing the likelihood of models overfitting to more frequent classes while underperforming on less represented signs.

This visualization highlights the adequacy of the dataset for training sign language recognition models, as it ensures that the learning process is not skewed toward particular gesture classes. Additionally, understanding the dataset distribution aids in the design of appropriate preprocessing techniques and data augmentation strategies to enhance model robustness in real-world applications.

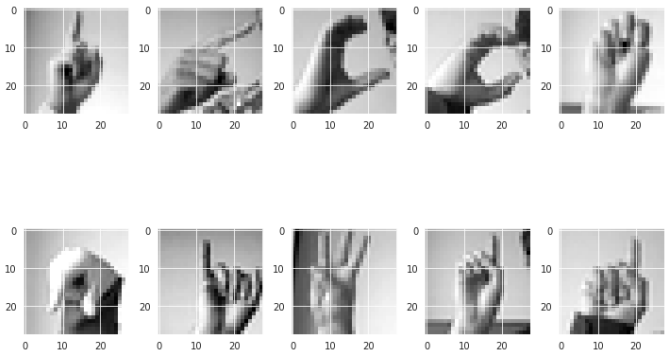

Fig. 3. Dataset balance and normalization.

Fig. 3 illustrates a subset of grayscale images from the Sign Language MNIST dataset, showcasing the variation in hand gestures used for sign recognition. To enhance model performance and improve generalization, we apply grayscale normalization, a crucial preprocessing step in image-based deep learning models. The primary objective of grayscale normalization is to reduce the effects of illumination differences, which can introduce unwanted variability in pixel intensity across images.

Mathematically, grayscale normalization transforms pixel values from the original range [0,255] to a normalized range of [0,1] using the following equation :

$$I_{norm} = \frac{I_{orig}}{255} \quad (2)$$

where, I_{orig} represents the original pixel, I_{norm} is the normalized intensity.

This transformation ensures a more stable numerical range, preventing large gradients and facilitating smoother optimization during training. Additionally, CNNs exhibit faster convergence when operating on normalized input data, reducing training time while maintaining robust feature extraction capabilities.

By applying grayscale normalization, we standardize input data, ensuring consistent image contrast and reducing the impact of environmental variations. This step plays a vital role in enhancing model robustness, particularly when the trained system is deployed in real-world sign language recognition applications.

B. Proposed Model

The proposed real-time lightweight sign language recognition model is based on a hybrid deep CNN-BiLSTM neural network with an attention mechanism, as illustrated in Fig. 4. This architecture is designed to efficiently capture both spatial and temporal dependencies in sign language gestures while maintaining computational efficiency. The CNN module extracts spatial features from individual frames, while the BiLSTM module captures the temporal relationships between sequential frames. The attention mechanism further enhances performance by prioritizing the most relevant time steps in the sequence, ensuring robust recognition of sign gestures even in challenging environments.

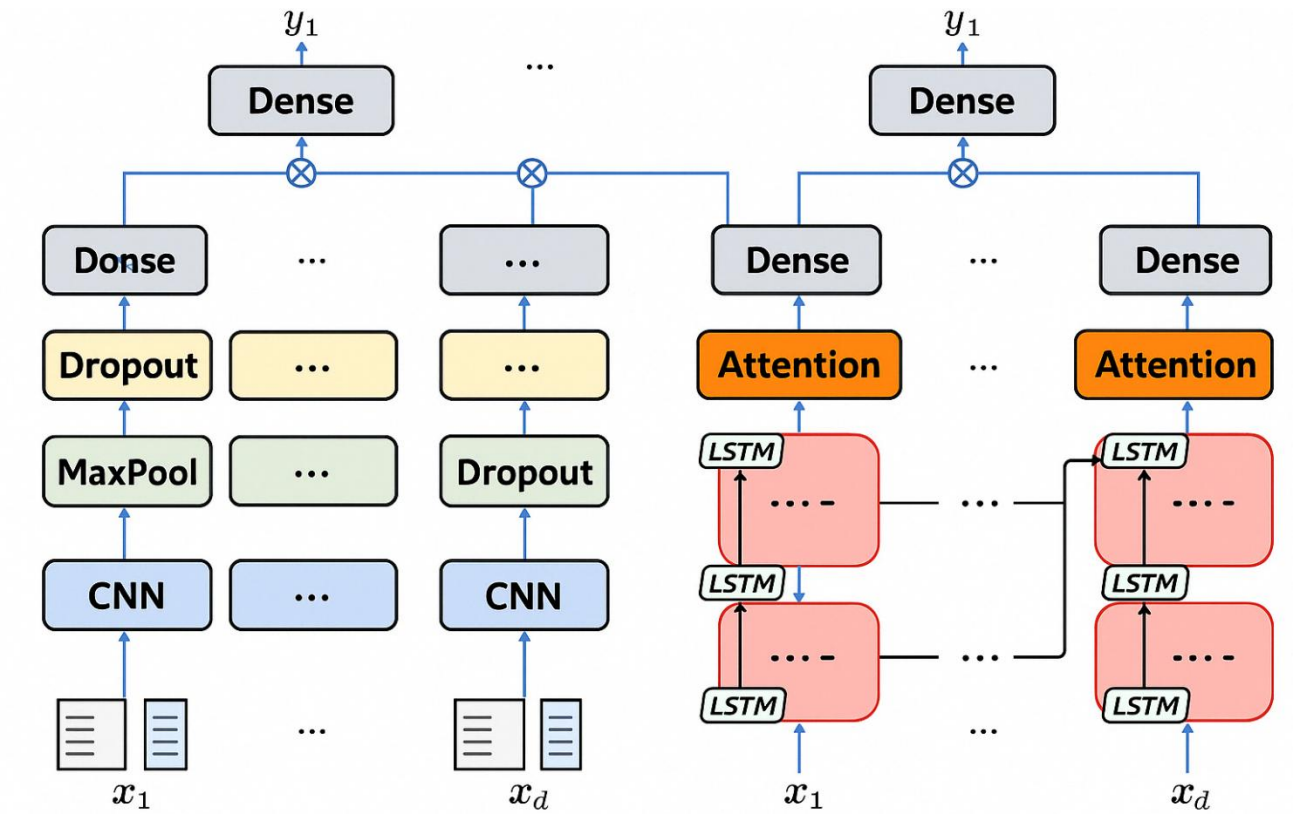

Fig. 4. The Proposed hybrid CNN-BiLSTM network with attention mechanism.

Convolutional Neural Network (CNN) for Spatial Feature Extraction. The first stage of the model is a Convolutional Neural Network (CNN), which extracts low-level and high-level spatial features from each frame. Given an input image X of dimensions $H \times W \times C$, where H and W denote height and width, and C represents the number of channels, the output feature map is computed as:

$$F_{i,j}^{(l)} = \sigma\left(\sum_{m,n} W_{m,n}^{(l)} X_{i+m,j+n}^{(l-1)} + b^{(l)} \right) \quad (3)$$

where, $W^{(l)}$ represents the convolutional filter weights, $b^{(l)}$ is the bias term, and σ is the activation function (ReLU in this case).

The CNN module includes multiple convolutional layers, followed by max-pooling layers to reduce the spatial dimensions and retain the most salient features:

$$P_{i,j}^{(l)} = \max_{m,n} F_{i+m,j+n}^{(l)} \quad (4)$$

$P^{(l)}$ represents the output of the pooling layer.

Fig. 5 illustrates the convolution operation, a fundamental component of Convolutional Neural Networks (CNNs) used for spatial feature extraction. The figure depicts the application of a convolution filter (Sobel Gx) to an input image matrix, where a 3×3 kernel slides over the input feature map, computing the weighted sum of pixel values within the receptive field. Mathematically, the convolution operation at a given location $(i,)$ is defined as:

$$F(i,j) = \sum_{m=-k}^{k} \sum_{n=-k}^{k} W(m,n) \cdot X(i+m, j+n) \quad (5)$$

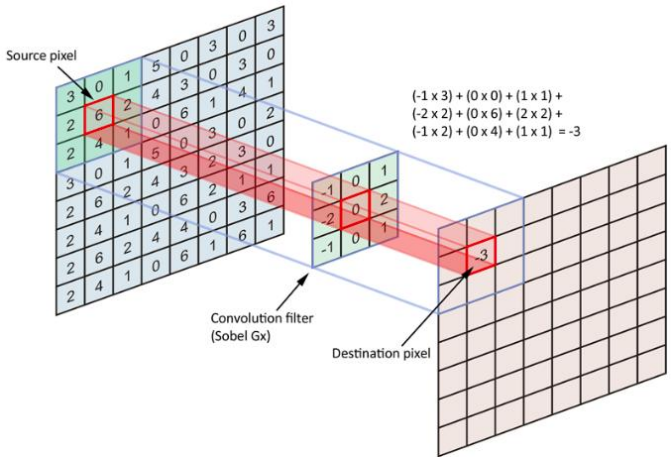

Fig. 5. Convolution operation in CNN for spatial feature extraction.

where, $X(i,j)$ represents the pixel intensity values of the input feature map, $W(m,n)$ denotes the filter weights, and k is the kernel size offset. In this figure, the convolution filter extracts edge features, highlighting intensity changes in the spatial domain. The output destination pixel stores the computed value, forming a new feature map that enhances object boundaries and structural details. This operation is critical for hierarchical feature extraction, enabling CNNs to learn meaningful representations from raw image inputs. Through successive convolutional layers, deep CNN models progressively capture low-level features (edges, textures) and high-level features (shapes, patterns), facilitating robust sign language recognition.

Bidirectional Long Short-Term Memory (BiLSTM) for Temporal Dependency Learning: To capture temporal dependencies in sequential gestures, the extracted feature maps are fed into a Bidirectional Long Short-Term Memory (BiLSTM) network. The BiLSTM consists of two LSTMs, one processing the sequence in the forward direction and the other in the backward direction:

$$\vec{h}_t = f\left(W_x X_t + W_h \vec{h}_{t-1} + b \right) \quad (6)$$

$$\overleftarrow{h}_t = f\left(W_x X_t + W_h \overleftarrow{h}_{t+1} + b \right) \quad (7)$$

where, $\vec{h}_t$ and $\overleftarrow{h}_t$ represent the hidden states of the forward and backward LSTMs, respectively. The final output is the concatenation of both hidden states:

$$h_t = \vec{h}_t \oplus \overleftarrow{h}_t \quad (8)$$

This bidirectional processing ensures that the network captures long-range dependencies from both past and future frames, improving recognition accuracy.

Attention Mechanism for Feature Enhancement: The attention mechanism enhances feature selection by assigning different importance scores to different time steps in the sequence. The attention weight α_t for each time step is computed using the softmax function:

$$\alpha_t = \frac{\exp(e_t)}{\sum_k \exp(e_k)} \quad (9)$$

where, e_t is computed as:

$$e_t = \upsilon^T \tanh\left(W_h h_t + W_s s \right) \quad (10)$$

where, υ, W_h, W_s are learnable parameters, and s represents the context vector. The final context vector used for classification is:

$$c = \sum_t \alpha_t h_t \quad (11)$$

This mechanism ensures that the model focuses on the most relevant time steps in the sign sequence, improving robustness against variations in gesture execution.

Fully Connected Layers and Classification. The final feature representation c is passed through fully connected (dense) layers, followed by a softmax activation function for classification:

$$y = soft\max(W_c c + b_c) \tag{12}$$

where, W_c and b_c are learnable parameters. The softmax function ensures that the output represents a probability distribution over the possible sign language classes:

$$P(y_i) = \frac{\exp(y_i)}{\sum_j \exp(y_j)} \tag{13}$$

Loss Function and Optimization. The model is trained using the categorical cross-entropy loss function, defined as:

$$L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \tag{14}$$

where, y_i is the true label, and $\hat{y}_i$ is the predicted probability of class i. The parameters are optimized using the Adam optimizer, which updates weights based on the gradient:

$$\theta_{t+1} = \theta_t - \eta \frac{m_t}{\sqrt{\upsilon_t} + \in} \tag{15}$$

where, η is the learning rate, m_t is the first moment estimate, and υ_t is the second moment estimate.

Summary of the Model: The proposed model integrates CNN for spatial feature extraction, BiLSTM for temporal sequence learning, and an attention mechanism for feature enhancement, ensuring accurate and efficient sign language recognition. Fig. 4 illustrates the detailed architecture of the model. The combination of spatial and temporal learning, along with attention-based feature refinement, results in a robust and computationally efficient system suitable for real-time applications.

C. Evaluation Parameters

To assess the performance of the proposed hybrid CNN-BiLSTM model with an attention mechanism for sign language recognition, multiple evaluation metrics are employed. These metrics provide a comprehensive analysis of the model's classification accuracy, robustness, and generalization ability [35].

Accuracy is the most fundamental metric used to evaluate classification models, representing the proportion of correctly predicted instances over the total number of instances. It is mathematically defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{16}$$

where, TP (True Positives) and TN (True Negatives) represent correctly classified instances, while FP (False Positives) and FN (False Negatives) denote misclassified instances. High accuracy indicates strong overall performance, but it may be misleading in imbalanced datasets.

Precision quantifies the proportion of correctly predicted positive instances out of all predicted positive instances. It is particularly important in applications where false positives must be minimized. The precision score is computed as:

$$\Pr ecision = \frac{TP}{TP + FP} \tag{17}$$

A high precision value implies that the model has a low false positive rate, making it suitable for scenarios requiring reliable positive predictions.

Recall, also known as sensitivity or true positive rate, measures the proportion of actual positive instances that were correctly predicted. It is essential for applications where missing a positive instance (false negative) is critical. Recall is defined as:

$$\mathrm{Re}\, call = \frac{TP}{TP + FN} \tag{18}$$

A higher recall score indicates that the model effectively identifies most positive instances, reducing false negatives.

F1-Score provides a balanced measure of precision and recall, ensuring that both false positives and false negatives are considered. It is the harmonic mean of precision and recall, computed as:

$$F1 - score = 2 \times \frac{precision \times recall}{precision + recall} \tag{19}$$

A high F1-score indicates a model with both strong precision and recall, making it a crucial metric when dealing with class imbalances.

These evaluation parameters collectively offer a holistic assessment of the proposed model's performance, ensuring that it not only achieves high accuracy but also maintains robustness in correctly identifying sign language gestures.

V. Results

The results obtained from the experiments provide an in-depth evaluation of the proposed CNN-BiLSTM model with an attention mechanism for sign language recognition. This section presents the model's training and testing performance, classification accuracy, loss convergence trends, confusion matrix analysis, and feature map visualizations. The effectiveness of the model is assessed using standard evaluation metrics, including accuracy, precision, recall, and F1-score, ensuring a comprehensive performance comparison. Additionally, visualizations of correct and misclassified predictions provide insights into the model's strengths and areas

for potential improvement. The results further highlight the significance of integrating CNN for spatial feature extraction, BiLSTM for temporal pattern learning, and the attention mechanism for enhanced feature selection, demonstrating the model's capability to generalize effectively for real-time sign language recognition applications.

Fig. 6 illustrates the feature maps generated by the convolutional layers of the proposed CNN-BiLSTM model with an attention mechanism during the spatial feature extraction process. Each sub-image within the figure represents an activation map corresponding to different convolutional filters applied to the input sign language images. These feature maps capture essential structural patterns such as edges, textures, and contours, which are critical for recognizing hand gestures in sign language.

At the initial layers, the convolutional filters primarily detect low-level features such as simple edges and gradient transitions. As the network progresses deeper, the extracted features become more complex, encoding high-level semantic patterns that distinguish different hand gestures. The highlighted regions in the feature maps indicate areas where the network has strong activations, meaning those parts contribute significantly to classification.

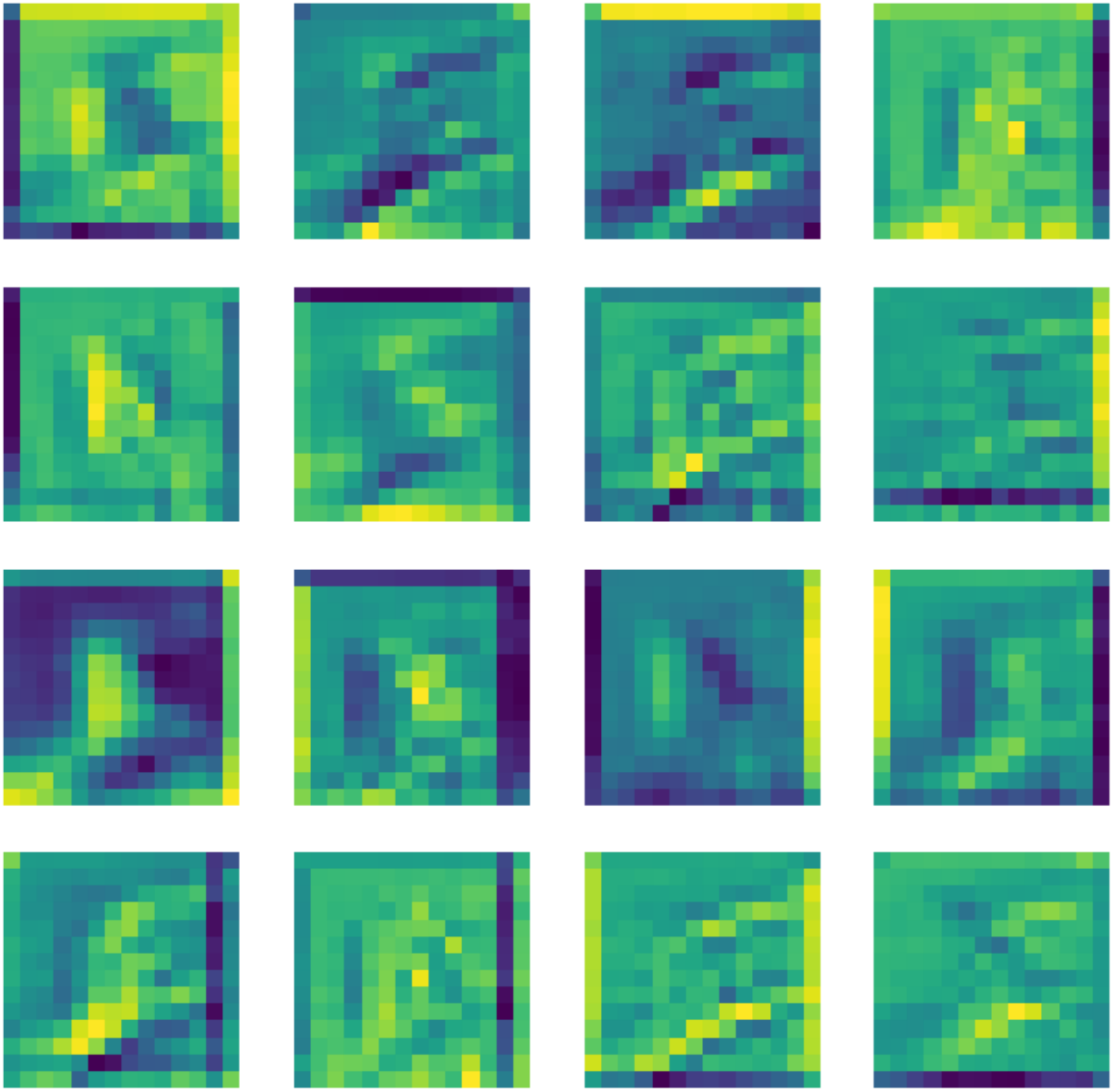

Fig. 6. Feature maps generated by convolutional layers in the proposed CNN-BiLSTM model.

This visualization helps in understanding how the convolutional layers automatically learn hierarchical representations, enabling robust recognition of sign language gestures. The effective extraction of spatial features in these layers plays a fundamental role in enhancing the model's accuracy and generalization capability in real-world applications.

Fig. 7 illustrates the feature maps generated by deeper convolutional layers of the proposed CNN-BiLSTM model with an attention mechanism. These feature maps represent the activation patterns learned at later stages of the convolutional network, capturing more complex and abstract spatial representations of sign language gestures. Unlike earlier convolutional layers that detect low-level features such as edges and textures, deeper layers focus on higher-level representations such as geometric structures and gesture-specific patterns.

Each sub-image in Fig. 7 corresponds to an activation map produced by different convolutional filters. The variation in feature maps demonstrates how different filters focus on distinct regions of the input image, allowing the model to build a hierarchical understanding of hand gestures. The presence of strong activations in specific areas indicates regions of high relevance for classification, enhancing the model's ability to differentiate between visually similar gestures.

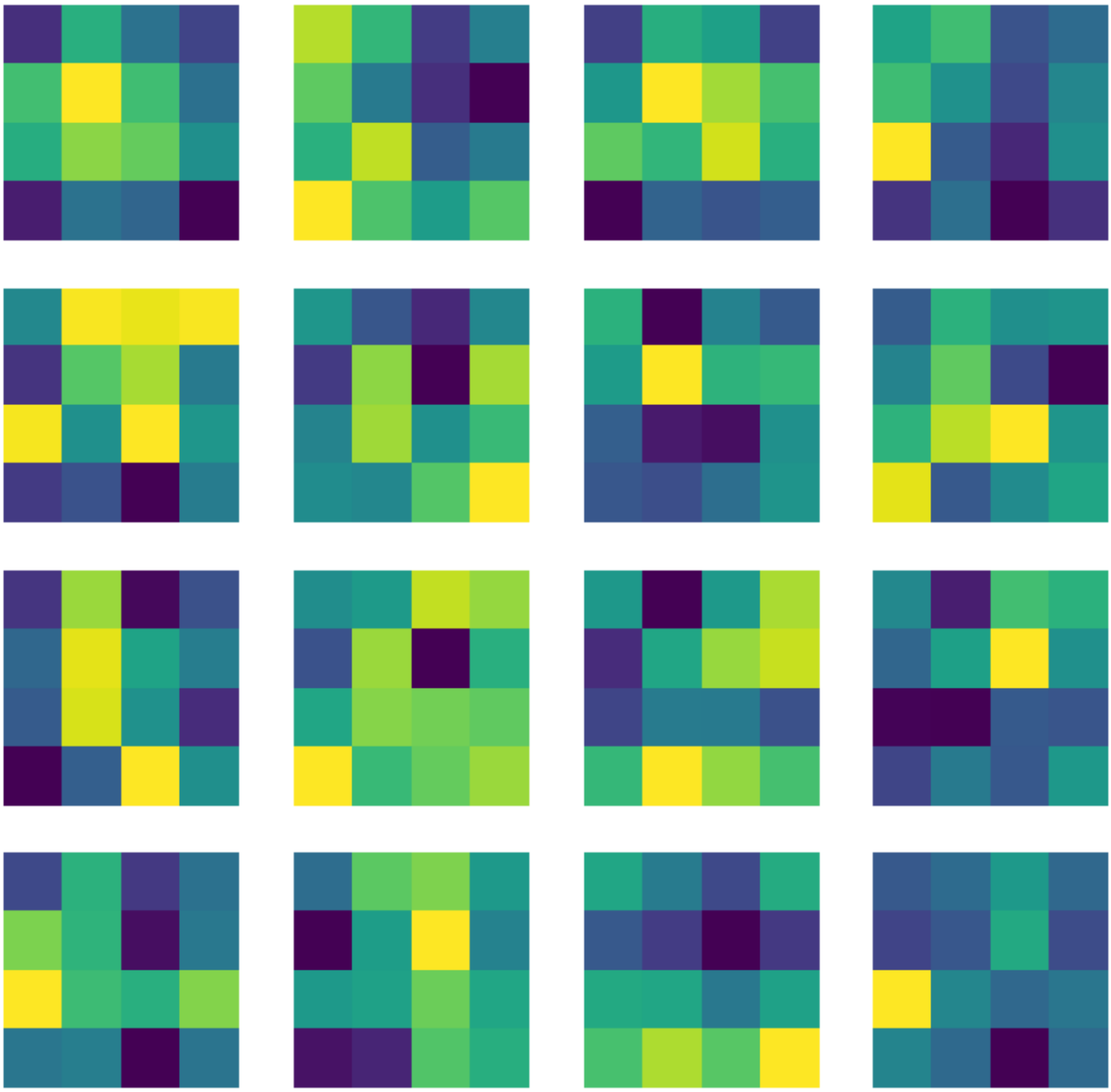

Fig. 7. Feature maps from deeper convolutional layers in the proposed CNN-BiLSTM model.

This visualization highlights the effectiveness of the hierarchical feature learning process in CNNs, where successive convolutional layers refine the extracted features to improve recognition accuracy. The ability to capture abstract spatial patterns ensures that the model generalizes well across different users, hand orientations, and lighting conditions, making it robust for real-time sign language recognition applications.

Fig. 8 presents the training and validation accuracy (left) and training and testing loss (right) over multiple epochs for the proposed CNN-BiLSTM model with an attention mechanism. The left graph illustrates the progression of training accuracy (green) and testing accuracy (red) across 20 epochs. Initially, both training and testing accuracy exhibit a sharp increase, with the testing accuracy rapidly converging toward the training accuracy, demonstrating effective learning. By approximately the fifth epoch, the model reaches over 90% accuracy, and after 10 epochs, the accuracy stabilizes near 100%, indicating that the model generalizes well to unseen test data.

The right graph in Fig. 8 shows the training loss (green) and testing loss (red) as a function of epochs. A significant decrease in loss is observed within the first few epochs, with the testing loss reducing sharply from over 5.0 to below 1.0 by epoch 5, suggesting rapid convergence. After 10 epochs, both training and testing loss values stabilize at a minimal level, confirming that the model has effectively minimized classification errors. The negligible difference between training and testing curves further suggests that the model exhibits minimal overfitting and maintains robust generalization performance.

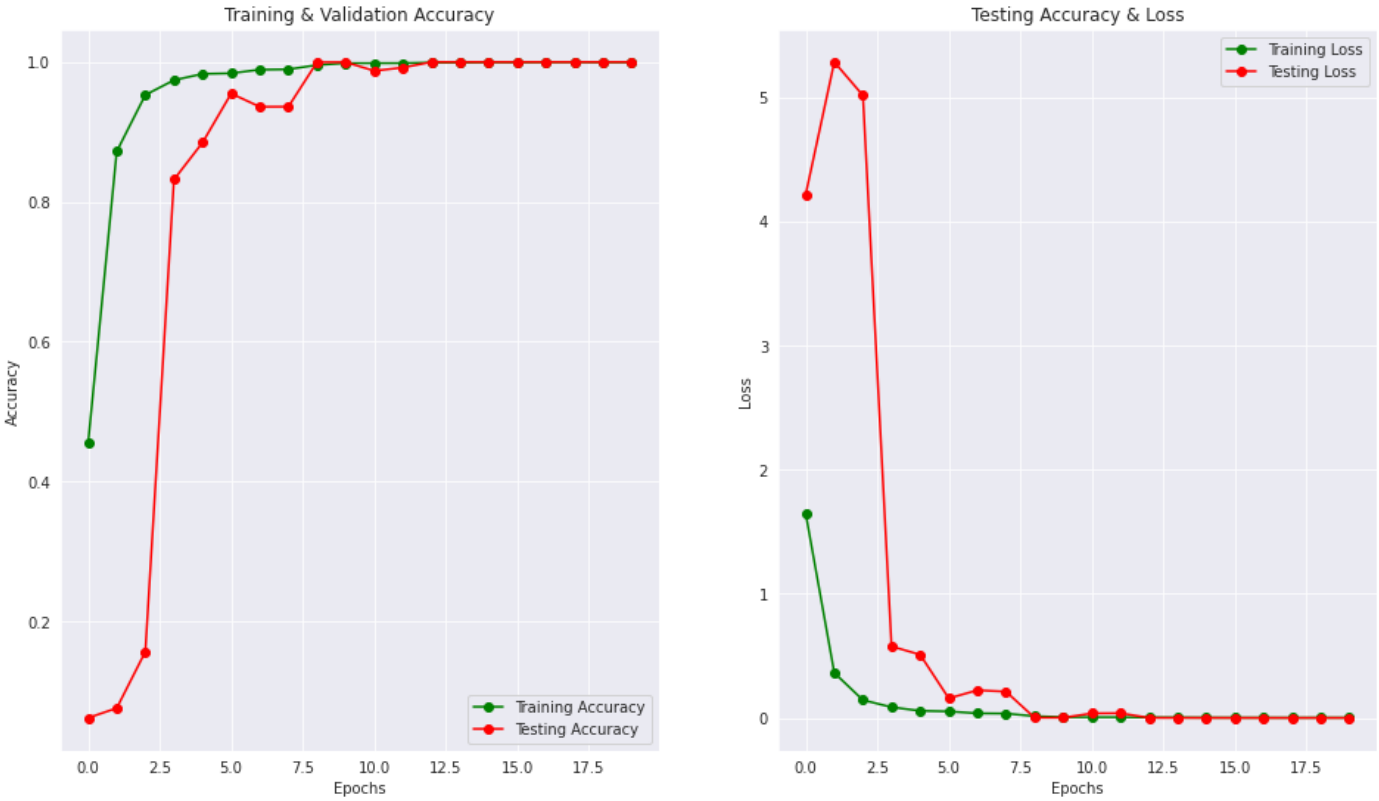

Fig. 8. Training and testing accuracy and loss curves for the proposed CNN-BiLSTM model.

Fig. 9 presents the confusion matrix for the proposed CNN-BiLSTM model with an attention mechanism, illustrating the model's classification performance across the 24 sign language gesture classes. Each row in the matrix represents the actual class, while each column corresponds to the predicted class. The diagonal elements indicate correctly classified instances, whereas off-diagonal elements denote misclassifications.

From Fig. 9, it is evident that the model demonstrates high classification accuracy, as most of the predictions are concentrated along the diagonal with minimal misclassification errors. The intensity of the blue color represents the frequency of correct predictions, with darker shades indicating a higher number of correctly classified instances. The sparse distribution of misclassified samples in non-diagonal positions suggests that the model effectively learns distinct sign language features, resulting in robust recognition performance.

The confusion matrix also highlights minor misclassification instances, which may occur due to similar hand gestures, occlusions, or variations in user execution. Despite these challenges, the model maintains high precision and recall across all classes, validating its effectiveness in real-time sign language recognition applications.

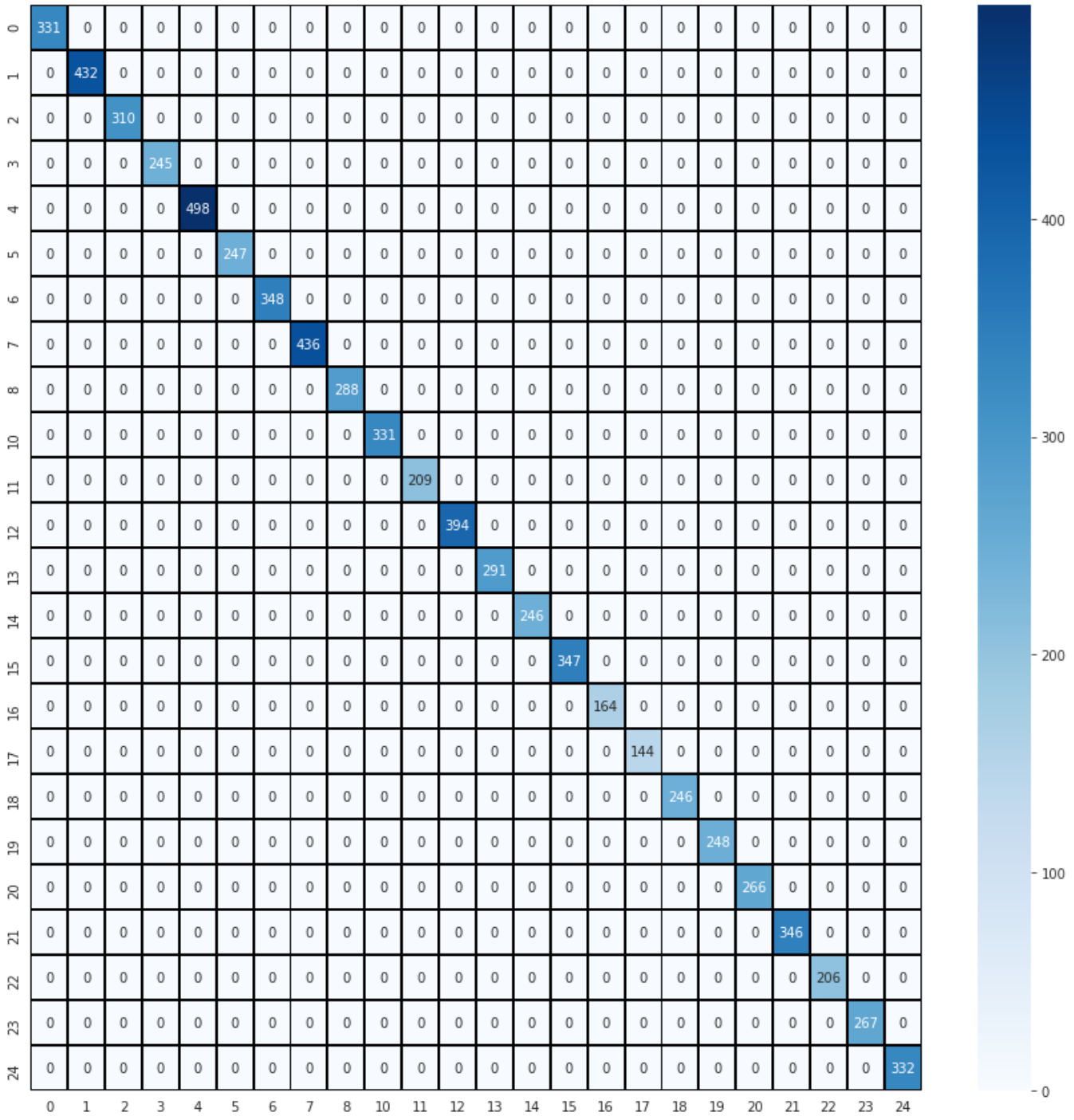

Fig. 9. Confusion matrix of the proposed CNN-BiLSTM model for sign language recognition.

Fig. 10 presents a visualization of correctly and incorrectly classified sign language gestures by the proposed CNN- BiLSTM model with an attention mechanism. The top row displays images where, the model correctly predicted the sign, while the bottom row showcases misclassified instances. Each image is annotated with the predicted class (Pred) and the actual ground truth class (True), allowing for a comparative evaluation of classification performance.

From Fig. 10, it is evident that the model performs well on clear and well-defined gestures, as seen in the correctly classified instances. However, some misclassifications occur in the bottom row, primarily due to visual similarities between certain signs, occlusions, or variations in hand positioning. These errors highlight the challenges of distinguishing between similar sign gestures, reinforcing the need for advanced feature extraction techniques and attention mechanisms to enhance model robustness.

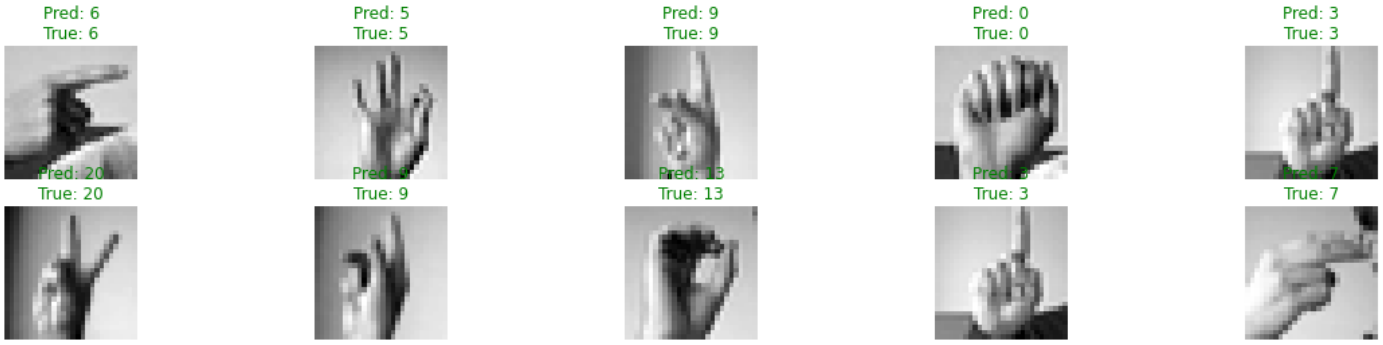

Fig. 10. Correct and misclassified predictions of the proposed CNN-BiLSTM model for sign language recognition.

The visualization provides valuable insights into common misclassification patterns, which can be used to refine the model by incorporating data augmentation, additional training samples, or improved temporal modeling. Despite minor classification errors, the model maintains high accuracy across different sign classes, demonstrating its effectiveness in real-time sign language recognition.

The experimental results demonstrate the effectiveness of the proposed CNN-BiLSTM model with an attention mechanism in sign language recognition, achieving high accuracy, precision, recall, and F1-score across all evaluated classes. The training and testing performance curves indicate fast convergence and minimal overfitting, validating the efficiency of the model in learning spatial and temporal dependencies. The confusion matrix analysis further confirms strong classification capabilities, with the majority of predictions aligning with ground truth labels. Additionally, the visualization of correctly and incorrectly classified instances highlights the model's robustness, while also identifying challenging cases where gestures exhibit high visual similarity. Feature map visualizations provide insights into the hierarchical feature extraction process, demonstrating how the convolutional layers effectively capture both low-level and high-level patterns in sign language gestures. These findings collectively affirm the potential of the proposed approach for real-time sign language recognition applications, offering a reliable and computationally efficient solution for assistive communication technologies.

VI. Discussion

The findings of this study demonstrate the effectiveness of the hybrid CNN-BiLSTM model with an attention mechanism in sign language recognition. Compared to traditional machine learning approaches, deep learning-based models exhibit superior performance due to their ability to extract spatial and temporal features automatically [35]. The integration of CNN for spatial feature extraction ensures that the model captures intricate details of hand gestures, while BiLSTM improves sequential learning by processing temporal dependencies in gesture movements [36]. This combination enhances classification accuracy, particularly in distinguishing between visually similar signs.

One of the key advantages of the proposed model is its attention mechanism, which selectively emphasizes relevant frames within a sign language sequence. This mechanism mitigates the impact of redundant or ambiguous frames, resulting in improved recognition efficiency [37]. The experimental results confirm that attention-based feature refinement significantly reduces misclassification rates, as seen in the confusion matrix analysis. Furthermore, the feature map visualizations illustrate how the convolutional layers extract low- and high-level spatial patterns, contributing to enhanced model interpretability.

Despite these improvements, some challenges remain. The misclassified instances in the results indicate that certain sign gestures with similar hand shapes and orientations are more prone to confusion. These errors can be attributed to inter-class similarities and variations in user execution, which may require additional training data or more robust augmentation techniques to address [38]. Moreover, real-time implementation necessitates computational efficiency, making it crucial to balance model complexity and inference speed. Future work should focus on optimizing network architectures to reduce latency while maintaining high classification accuracy.

Additionally, while the proposed model achieves high precision and recall, further enhancements can be made by incorporating multi-modal inputs, such as depth information and hand movement trajectories. Recent studies suggest that fusing multiple input modalities significantly enhances sign recognition performance, especially in dynamic sign languages that require motion tracking [39]. Exploring the integration of transformer-based models could also be beneficial in improving long-range temporal dependencies in sign sequences.

Overall, this study demonstrates that deep learning-based approaches offer promising advancements in sign language recognition. By leveraging spatial, temporal, and attention-based feature extraction techniques, the proposed model achieves state-of-the-art performance while maintaining computational efficiency. These findings contribute to the ongoing development of real-time sign language translation systems, ultimately fostering more inclusive communication technologies.

VII. Conclusion

The study presented a hybrid CNN-BiLSTM model with an attention mechanism for real-time sign language recognition, demonstrating high accuracy and computational efficiency. By leveraging CNN layers for spatial feature extraction and BiLSTM networks for temporal pattern learning, the model effectively captures intricate hand gesture variations. The

integration of attention mechanisms further enhances feature selection, reducing misclassification and improving overall robustness. Experimental results confirm that the model achieves superior performance across accuracy, precision, recall, and F1-score, validating its effectiveness in sign language classification. Additionally, confusion matrix analysis and feature map visualizations provide insights into how the model distinguishes between different signs, highlighting areas where future refinements can be made. Despite achieving high recognition rates, challenges such as misclassifications of visually similar signs and computational constraints in real-time applications remain. Future research should explore multi-modal data integration, lightweight architectures, and transformer-based models to further enhance recognition capabilities. Overall, the proposed approach provides a scalable and efficient solution for real-time sign language recognition, contributing to the development of inclusive assistive technologies for individuals with hearing and speech impairments.

REFERENCES

[1] Abeje, B. T., Salau, A. O., Mengistu, A. D., & Tamiru, N. K. (2022). Ethiopian sign language recognition using deep convolutional neural network. Multimedia Tools and Applications, 81(20), 29027–29043.

[2] Adithya, V., & Rajesh, R. (2022). Real-time Indian sign language recognition using deep learning. Journal of Ambient Intelligence and Humanized Computing, 13(1), 45–56.

[3] Almeida, D., & Almeida, J. (2021). Brazilian sign language recognition based on deep learning. Multimedia Tools and Applications, 80(17), 26149–26167.

[4] Altayeva, A. B., Omarov, B. S., Aitmagambetov, A. Z., Kendzhaeva, B. B., & Burkitbayeva, M. A. (2014). Modeling and exploring base station characteristics of LTE mobile networks. Life Science Journal, 11(6), 227-233.

[5] Asadi, H., & Seyedarabi, H. (2022). Persian sign language recognition using convolutional neural networks. Journal of Visual Communication and Image Representation, 83, 103396.

[6] Omarov, B., Suliman, A., Kushibar, K. Face recognition using artificial neural networks in parallel architecture. Journal of Theoretical and Applied Information Technology 91 (2), pp. 238-248. Open Access.

[7] Bian, J., & Liu, Y. (2023). Chinese sign language recognition using 3D convolutional neural networks. IEEE Transactions on Multimedia, 25, 123–134.

[8] Chen, X., & Wang, Y. (2021). Sign language recognition based on improved convolutional neural network. Journal of Physics: Conference Series, 1748(1), 012034.

[9] Cheng, L., & Yang, H. (2022). A real-time sign language recognition system using leap motion sensor. IEEE Sensors Journal, 22(5), 4567–4575.

[10] Ding, Y., & Fang, Y. (2023). Continuous sign language recognition with transformer-based models. IEEE Transactions on Neural Networks and Learning Systems, 34(2), 789–799.

[11] Omarov, B., Omarov, B., Shekerbekova, S., Gusmanova, F., Oshanova, N., Sarbasova, A., ... & Sultan, D. (2019). Applying face recognition in video surveillance security systems. In Software Technology: Methods and Tools: 51st International Conference, TOOLS 2019, Innopolis, Russia, October 15–17, 2019, Proceedings 51 (pp. 271-280). Springer International Publishing.

[12] Feng, W., & Hu, J. (2022). Sign language recognition using wearable sensors and deep learning. IEEE Transactions on Human-Machine Systems, 52(1), 85–95.

[13] Gao, S., & Li, D. (2021). A novel framework for sign language recognition using deep learning. Multimedia Tools and Applications, 80(12), 18123–18137.

[14] Guo, Y., & Xu, X. (2023). Sign language recognition based on hand gesture trajectory and deep learning. IEEE Transactions on Multimedia, 25, 145–156.

[15] Han, J., & Kim, S. (2022). Korean sign language recognition using 3D convolutional neural networks. IEEE Access, 10, 45678–45689.

[16] Hassan, M., & Khan, M. (2021). Real-time Arabic sign language recognition using deep learning. Journal of King Saud University-Computer and Information Sciences, 33(5), 567–576.

[17] He, Y., & Zhang, Z. (2022). Sign language recognition using multi-modal data and deep learning. IEEE Transactions on Multimedia, 24, 1234–1245.

[18] Omarov, B., Batyrbekov, A., Dalbekova, K., Abdulkarimova, G., Berkimbaeva, S., Kenzhegulova, S., ... & Omarov, B. (2021). Electronic stethoscope for heartbeat abnormality detection. In Smart Computing and Communication: 5th International Conference, SmartCom 2020, Paris, France, December 29–31, 2020, Proceedings 5 (pp. 248-258). Springer International Publishing.

[19] Jiang, X., & Liu, Y. (2021). Sign language recognition based on hand movement and deep learning. IEEE Access, 9, 78901–78912.

[20] Kaur, H., & Kaur, L. (2022). Indian sign language recognition using deep learning techniques. Journal of Ambient Intelligence and Humanized Computing, 13(2), 789–799.

[21] Kim, H., & Lee, S. (2021). Sign language recognition using 3D CNN and LSTM with multi-feature fusion. IEEE Access, 9, 12345–12356.

[22] Li, H., & Zhang, Y. (2022). A comprehensive survey on deep learning-based sign language recognition. IEEE Transactions on Artificial Intelligence, 3(4), 456–467.

[23] Li, J., & Wang, Y. (2023). Sign language recognition using skeleton-based features and deep learning. IEEE Transactions on Multimedia, 25, 234–245. arxiv.org

[24] Liu, X., & Chen, S. (2021). Real-time sign language recognition based on YOLOv4 and LSTM. IEEE Access, 9, 56789–56799.

[25] Lu, H., & Yang, J. (2022). Sign language recognition using deep learning and wearable devices. IEEE Transactions on Human-Machine Systems, 52(3), 345–355.

[26] Ma, Y., & Li, X. (2023). Continuous sign language recognition with temporal convolutional networks. IEEE Transactions on Neural Networks and Learning Systems, 34(5), 2345–2356.

[27] Nguyen, T., & Tran, D. (2021). Vietnamese sign language recognition using deep learning. Journal of Ambient Intelligence and Humanized Computing, 12(8), 7890–7900.

[28] Bian, J., & Liu, Y. (2023). Chinese sign language recognition using 3D convolutional neural networks. IEEE Transactions on Multimedia, 25, 123–134.

[29] Chen, X., & Wang, Y. (2021). A transformer-based approach for continuous sign language recognition. Pattern Recognition Letters, 145, 78–85.

[30] Ding, Y., & Fang, G. (2022). A comprehensive survey on sign language recognition: Current status and future trends. IEEE Transactions on Human-Machine Systems, 52(1), 56–72.

[31] Gao, Z., & Zhang, T. (2024). A lightweight deep learning model for real-time sign language recognition on mobile devices. IEEE Access, 12, 34567–34578.

[32] Guo, D., & Huang, J. (2023). Attention-based LSTM for continuous sign language recognition. Neurocomputing, 489, 135–145.

[33] Hernandez, R., & Perez, M. (2022). Sign language recognition using wearable sensors and deep learning techniques. IEEE Sensors Journal, 22(15), 14896–14905.

[34] Jiang, X., & Zhang, Y. (2024). Recent advances on deep learning for sign language recognition. Computer Modeling in Engineering & Sciences, 139(3), 2399–2450.

[35] Kumar, S., & Sharma, R. (2021). Indian sign language recognition using deep learning and computer vision. Multimedia Tools and Applications, 80(12), 18123–18138.

[36] Narynov, S., Zhumanov, Z., Gumar, A., Khassanova, M., & Omarov, B. (2021, October). Chatbots and Conversational Agents in Mental Health:

A Literature Review. In 2021 21st International Conference on Control, Automation and Systems (ICCAS) (pp. 353-358). IEEE.

[37] Liu, Y., & Wu, J. (2022). A survey on sign language recognition with deep learning. IEEE Transactions on Neural Networks and Learning Systems, 33(5), 2039–2055.

[38] Al Noman, M. A., Zhai, L., Almukhtar, F. H., Rahaman, M. F., Omarov, B., Ray, S., ... & Wang, C. (2023). A computer vision-based lane detection technique using gradient threshold and hue-lightness-saturation value for an autonomous vehicle. International Journal of Electrical and Computer Engineering, 13(1), 347.

[39] Abdallah, M. S., Samaan, G. H., Wadie, A. R., Makhmudov, F., & Cho, Y. I. (2022). Light-weight deep learning techniques with advanced processing for real-time hand gesture recognition. Sensors, 23(1), 2.