



Real-Time Indonesian Sign Language Recognition Using Long Short-Term Memory and IndoBERT Contextual Modeling

Amanda Betania Maritza^{1*}, Rizka Ardiansyah²

¹Informatics Engineering Study Program, Department of Information Technology,
Faculty of Engineering, Universitas Tadulako, Indonesia

²Department of Information Technology, Faculty of Engineering,
Universitas Tadulako, Indonesia

E-Mail: ¹f55122032@untad.ac.id, ²rizka@untad.ac.id

Received Apr 26th 2026; Revised May 29th 2026; Accepted Jun 17th 2026; Available Online Jun 27th 2026

Corresponding Author: Amanda Betania Maritza

Copyright © 2026 by Authors, Published by Institut Riset dan Publikasi Indonesia (IRPI)

Abstract

Communication barriers between deaf and hearing individuals remain a major challenge in inclusive education and social interaction. Most Indonesian Sign Language (*Sistem Isyarat Bahasa Indonesia* = SIBI) recognition systems focus solely on isolated gesture classification, without contextual understanding. This study proposes a real-time SIBI recognition system integrating LSTM-based temporal modeling and IndoBERT contextual language modeling. Hand keypoints were extracted from gesture sequences and processed using Long Short-Term Memory (LSTM) to recognize dynamic alphabet gestures. The predicted letter sequences were refined using a Trie-based lexical filter and IndoBERT to generate contextually appropriate word predictions. The dataset consisted of 780 gesture sequences representing 26 alphabet classes. The results showed that the system achieved 89.42% accuracy, 91.57% precision, 89.42% recall, and 88.74% F1-score while maintaining real-time performance at an average processing speed of 249.18 FPS and a latency of 129.46 ms per sample. Statistical evaluation confirmed that camera distance significantly affected recognition performance, while integrating IndoBERT improved contextual word-prediction accuracy from 18.2% to 58.64% (3.2× improvement). McNemar analysis further verified that the improvement was statistically significant, highlighting the effectiveness of contextual language modeling for enhancing semantic interpretation in SIBI recognition. The proposed framework improves semantic interpretation and real-time usability for SIBI-based assistive communication systems.

Keywords: Contextual Language Modeling, IndoBERT, LSTM, Real-Time Recognition, SIBI

1. INTRODUCTION

Communication plays an important role in education, collaboration, and social interaction. However, individuals with hearing impairments still experience significant barriers in accessing information and communicating effectively with hearing communities [1], [2]. In Indonesia, *Sistem Isyarat Bahasa Indonesia* (SIBI) is widely used in formal educational environments and institutional communication. However, limited public understanding of SIBI continues to create communication gaps, particularly in higher education environments where interaction and information exchange are essential [3]. Along with the growing adoption of Artificial Intelligence (AI) in educational technology, intelligent assistive systems have become a promising solution to improve accessibility and inclusivity for deaf individuals [4], [5], [6]. Despite these developments, the availability of accurate and real-time SIBI recognition systems that can operate efficiently on common hardware remains limited.

Sign language recognition (SLR) has attracted significant attention in recent years due to its potential to improve communication accessibility for deaf individuals. Various machine learning and deep learning techniques have been explored to recognize sign language gestures from visual input data. Early studies commonly used traditional classification methods such as K-Nearest Neighbor, Random Forest, and Support Vector Machine to recognize static hand gestures [7], [8]. However, these methods were incapable of modeling temporal characteristics of dynamic sign language movements. In order to extract spatial features, several studies have adopted Convolutional Neural Network (CNN)-based architectures such as Residual Network (ResNet), InceptionV3, and You Only Look Once (YOLO) [9], [10]. These approaches demonstrated improved visual representation capabilities but often required high computational resources and were less efficient for real-time implementation.

Modeling temporal dependencies in dynamic sign language gestures remains a major challenge in sign language recognition tasks. To address this issue, sequential deep learning architectures such as Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) have been extensively utilized for sign language recognition [11], [12]. LSTM-based approaches are capable of capturing sequential motion patterns effectively and have shown competitive performance in SIBI recognition studies [13], [14], [15]. Furthermore, the integration of MediaPipe keypoint extraction has significantly improved computational efficiency by representing gestures as skeletal landmark coordinates instead of raw image frames [13], [16]. This representation reduces input dimensionality while preserving important motion information, enabling lightweight real-time recognition systems.

Despite high classification accuracy in isolated gesture recognition reported in previous studies, most existing systems only predict the alphabet level or gesture level without context [17], [18], [19], [20], [21]. In real-life communication scenarios, isolated character recognition alone is insufficient because sign language interpretation requires semantic consistency and contextual word formation. As a result, many systems generate fragmented or semantically inconsistent outputs despite achieving high gesture classification performance. Recent advances in contextual language modeling, particularly Bidirectional Encoder Representations from Transformers (BERT)-based architectures, have demonstrated strong capability in capturing semantic relationships and contextual dependencies in Indonesian language tasks [22], [23]. Indonesia Bidirectional Encoder Representations from Transformers (IndoBERT) is especially suitable for this purpose because it was pretrained on large-scale Indonesian corpora, enabling better contextual understanding than general multilingual models.

Although previous studies have demonstrated promising results for SIBI recognition, several limitations remain. Many existing approaches focus primarily on improving gesture classification accuracy without evaluating semantic consistency in real-world communication scenarios. Furthermore, studies that employ LSTM and MediaPipe-based architectures generally emphasize character-level recognition and rarely investigate the integration of contextual language models to refine recognized letter sequences into meaningful words. In addition, comprehensive evaluations involving real-time performance, environmental robustness, and statistical validation of contextual prediction methods are still limited. These limitations highlight the need for a framework that combines temporal gesture recognition with contextual language understanding to support more accurate and semantically coherent SIBI interpretation.

To address the limitations, this study proposes a real-time Indonesian Sign Language recognition framework that integrates temporal gesture recognition and contextual language modeling. The proposed system combines MediaPipe-based hand keypoint extraction and LSTM-based temporal modeling for dynamic gesture recognition, followed by lexical filtering using a Trie structure and contextual refinement using IndoBERT. IndoBERT was chosen because it was pretrained on Indonesian datasets and provides better contextual understanding for Indonesian language tasks compared to general multilingual models. Through this integration, the framework is expected to improve not only gesture recognition accuracy but also semantic coherence in word prediction. This research aims to develop and evaluate a real-time SIBI recognition system that achieves high recognition performance while maintaining computational efficiency suitable for real-time implementation. The integration of temporal gesture modeling and contextual language understanding is expected to improve semantic coherence in word prediction and enhance real-time sign language interpretation. Experimental evaluation and statistical analysis were conducted to examine recognition performance, environmental robustness, and the effectiveness of contextual language modeling in improving semantic prediction accuracy. The proposed framework is expected to contribute to the development of intelligent assistive technologies that support inclusive communication and learning environments for deaf individuals.

2. MATERIALS AND METHOD

This study adopted an experimental approach to design and evaluate a real-time Indonesian Sign Language (SIBI) recognition framework. The research focuses on the implementation of deep learning-based temporal gesture recognition combined with contextual language modeling to support semantic interpretation in real-time communication scenarios. MediaPipe-based hand keypoint extraction and LSTM networks were utilized for dynamic gesture recognition, followed by Trie-based lexical filtering and IndoBERT contextual modeling for semantic word prediction.

The proposed framework consists of gesture data acquisition, preprocessing, temporal gesture modeling, contextual language prediction, and performance evaluation. The system was evaluated quantitatively using classification performance metrics, inference efficiency analysis, and statistical testing to examine system reliability and robustness under different testing conditions. The overall workflow of the framework is presented in Figure 1 (write image on Generative AI).

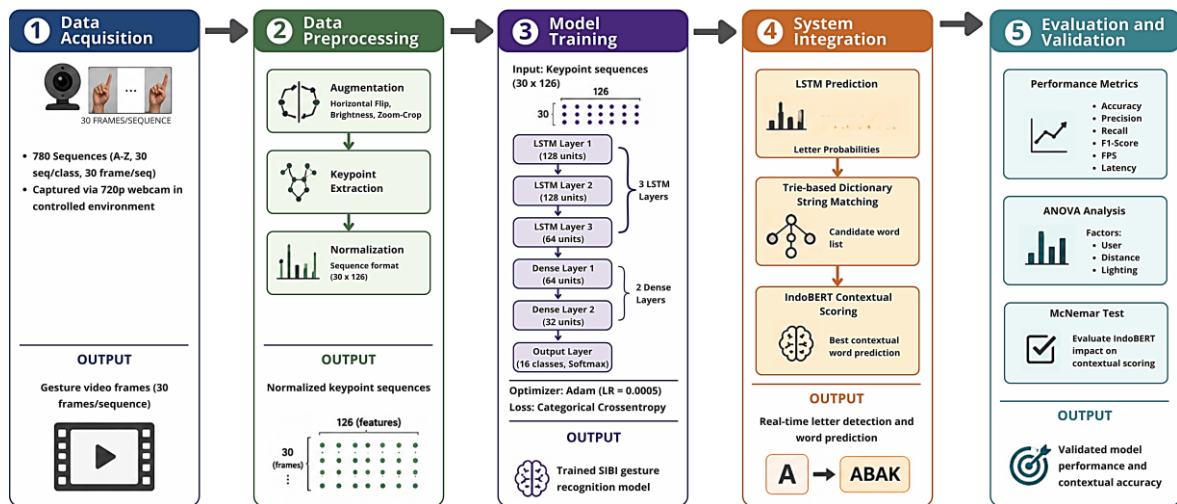


Figure 1. Research Workflow of the Real-Time SIBI Recognition System

2.1. Dataset Preparation and Data Collection

The dataset consisted of 780 gesture sequences representing 26 SIBI alphabet classes (A–Z). Each alphabet class contained 30 gesture sequences, where every sequence consisted of 30 consecutive frames representing one complete gesture movement. The gesture data were collected directly by the researcher using a built-in laptop webcam with a resolution of 720p at 30 FPS.

In addition, an inference testing dataset consisting of 78 video samples collected from five different participants (3×26 gestures) was prepared to evaluate real-time recognition performance. This dataset was used to assess the ability of the trained model to recognize gestures performed by unseen users with different signing characteristics.

The data collection process involved participants with diverse demographic and signing backgrounds to improve dataset representativeness. Participants included both male and female individuals aged between 16 and 40 years. Their SIBI experience varied, ranging from proficient SIBI users and students undergoing formal SIBI learning to individuals who performed gestures based on official SIBI reference materials. This diversity was intentionally incorporated to capture variations in hand shape, gesture execution style, and signing speed, thereby improving the representativeness of the dataset and supporting the development of a more robust recognition model.

Data acquisition was conducted in a controlled indoor environment with simple background and a close camera distance to minimize environmental noise and improve gesture visibility. During data acquisition, participants were instructed to perform each SIBI alphabet gesture individually while facing the camera. Each gesture sequence was recorded for 30 consecutive frames, and multiple repetitions were collected to capture natural variations in gesture execution.

The dataset was divided using a subject-wise split strategy consisting of 70% training data, 15% validation data, and 15% testing data. This separation ensured that gesture samples used during testing were not present in the training process, enabling reliable evaluation of the model generalization capability on unseen gesture sequences.

2.2. Data Preprocessing and Feature Extraction

Before model training, several preprocessing and augmentation techniques were applied to improve model robustness against environmental and positional variations. Each captured frame was augmented using horizontal flipping, brightness adjustment, and zoom-crop transformation. Brightness values were randomly adjusted within a range of 70% to 120% of the original image intensity to simulate different lighting conditions. In addition, a zoom factor of 1.2 was applied to generate variations in hand size and camera distance. These augmentation techniques were employed to increase data diversity and improve model robustness under different acquisition conditions.

Hand and body landmark extraction was performed using the MediaPipe framework [18], which detects landmarks from the hands, face, and upper body regions. This approach enabled stable localization of hand movements even under partial occlusion or slight rotational variation. However, only hand landmarks were utilized as model input because hand motion represents the primary information source in SIBI alphabet gestures.

Each frame produced 126 spatial coordinates consisting of x, y, and z values extracted from both hands. The extracted keypoints were normalized to reduce the influence of positional displacement, hand scale

variation, and illumination differences. The final processed data were structured into temporal sequences of size (30×126) , representing 30 frames with 126 normalized spatial features per sequence.

2.3. Model Architecture and Training Configuration

The proposed recognition model employed a sequential architecture based on stacked LSTM layers to capture temporal dependencies in sign language motion patterns. The overall layer configuration of the network is summarized in Table 1.

Table 1. Model Layer Configuration

Layer	Description
Input	(30, 126 features)
LSTM	128, 128, 64 units
Dense	64, 32 units
Dense	26 classes (A-Z), softmax activation

The input layer received sequential gesture data with dimensions of (30, 126). Three stacked LSTM layers containing 128, 128, and 64 hidden units were utilized to learn hierarchical temporal representations from dynamic hand movements. A dropout rate of 0.1 was applied after each LSTM layer to reduce overfitting and improve model generalization performance.

Following temporal feature extraction, two fully connected Dense layers with 64 and 32 neurons were employed for feature refinement and dimensionality reduction before classification. The final output layer consisted of 26 neurons corresponding to the SIBI alphabet classes (A–Z) using softmax activation to generate class probability distributions.

The model was trained for a maximum of 200 epochs using a batch size of 32, the Adam optimizer with a learning rate of 0.0005, and the categorical crossentropy loss function. To minimize overfitting during training, an early stopping mechanism with a patience threshold of 50 epochs was employed. In addition, learning rate adaptation was handled using the ReduceLROnPlateau strategy with a reduction factor of 0.5 to maintain stable convergence throughout the optimization process.

To improve evaluation reliability, a 10-fold cross-validation strategy was applied during training. The dataset was partitioned into ten equal subsets, where each iteration used nine subsets for model training and the remaining subset for validation. This approach was employed to reduce evaluation bias, improve model generalization assessment, and ensure the stability and consistency of the implemented LSTM architecture across different data partitions.

All experiments were conducted using TensorFlow 2.17.1 and Keras within the Visual Studio Code development environment. MediaPipe 0.10.13 was utilized for hand landmark extraction, OpenCV for image processing operations, and Scikit-learn for data partitioning and model evaluation. The experimental setup utilized an ASUS Vivobook K3402ZA laptop equipped with an Intel Core i7-12700H processor, 16 GB DDR4 RAM, and Intel Iris Xe Graphics.

2.4. System Integration and Contextual Prediction

The developed system integrated visual gesture recognition with contextual language prediction to generate semantically coherent text outputs in real time. The complete integration architecture is illustrated in Figure 2.

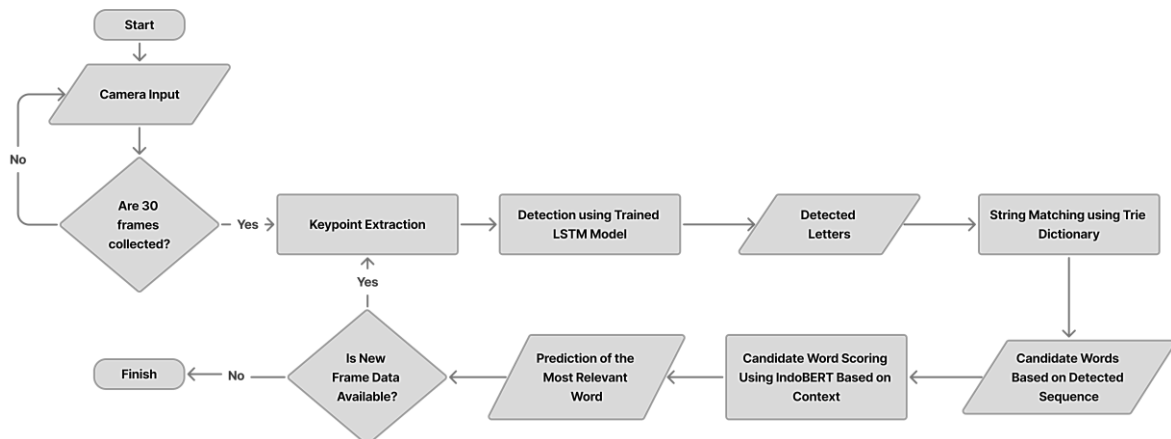


Figure 2. System Architecture of The Real-Time SIBI Recognition and Contextual Prediction Pipeline.

During system operation, the webcam continuously captured gesture sequences consisting of 30 frames. The extracted keypoint sequences were then processed by the trained LSTM model to predict alphabet labels based on temporal gesture dynamics.

The predicted character outputs were subsequently processed using a Trie-based dictionary structure to generate candidate words that matched predefined vocabulary entries. Contextual semantic refinement was then performed using IndoBERT embeddings, which assigned contextual probability scores to candidate words based on semantic relevance. The final predicted word was selected according to the maximum contextual probability score. This integration, combining deterministic lexical filtering and probabilistic contextual modeling, was designed to improve semantic consistency and reduce invalid word generation during continuous communication.

2.5. Experimental Design and Performance Evaluation

System performance was evaluated using accuracy, precision, recall, and F1-score, with macro averaging to ensure balanced evaluation across all gesture classes. Macro averaging computes evaluation metrics independently for each class and then averages them, preventing performance bias toward frequently predicted classes.

In the evaluation equations, True Positive (TP) represents correctly predicted gesture classes, True Negative (TN) represents correctly rejected non-target classes, False Positive (FP) represents incorrectly predicted gesture classes, and False Negative (FN) represents target gesture classes that were not correctly recognized by the system.

Accuracy represents the ratio of correctly recognized gesture samples to the total number of evaluated predictions. In this study, accuracy provides a global indicator of how reliably the system recognizes SIBI alphabet gestures across all testing conditions. The accuracy metric is calculated using Equation (1).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Precision indicates the percentage of predicted gesture classes that were correctly classified by the model. Precision reflects how often the model's predicted alphabet label is correct when it assigns that label. The precision value is computed using Equation (2).

$$Precision (P) = \frac{TP}{TP+FP} \quad (2)$$

Recall describes the ability of the model to successfully identify gesture classes that actually appeared in the testing data. In this research, recall indicates how effectively the model detects each alphabet gesture when it truly occurs. Recall is calculated using Equation (3).

$$Recall (R) = \frac{TP}{TP+FN} \quad (3)$$

F1-score combines precision and recall into a single metric using their harmonic mean, providing a balanced measurement of classification performance under both false-positive and false-negative conditions. The F1-score is determined using Equation (4).

$$F1 - Score = \frac{2(Precision \times Recall)}{Precision+Recall} \quad (4)$$

A confusion matrix was additionally generated to analyze classification behavior and identify gesture classes with overlapping spatial characteristics.

To evaluate the influence of environmental and user-related factors, a repeated-measures ANOVA was conducted using three independent variables, namely camera distance (near, medium, and far), lighting condition (bright and dim), and user variation involving two participants. Each participant performed all gesture classes under every experimental condition, forming a within-subject experimental design. Statistical analysis included F-values, p-values with significance level $\alpha = 0.05$, and effect size (η^2) to determine the significance and magnitude of each factor.

Furthermore, the contribution of IndoBERT contextual integration was analyzed using the McNemar statistical test based on paired prediction outcomes before and after contextual enhancement. The McNemar test statistic is presented in Equation (5).

$$\chi^2 = \frac{(|b-c|-1)^2}{b+c} \quad (5)$$

where b and c represent discordant prediction pairs. A p-value lower than 0.05 indicated that IndoBERT integration significantly improved contextual word prediction performance.

3. RESULTS AND DISCUSSION

3.1. Training Performance

The developed LSTM-based temporal modeling architecture was trained using the prepared gesture sequence dataset. Model performance during the training phase was evaluated using training and validation accuracy and loss metrics. The final training results are presented in Table 2.

Table 2. Model Training Performance

Metric	Value
Training Accuracy	99.71%
Training Loss	0.0190
Validation Accuracy	99.77%
Validation Loss	0.0189

Based on Table 2, the model achieved a training accuracy of 99.71% with a training loss of 0.0190, while validation accuracy reached 99.77% with a validation loss of 0.0189. The relatively high accuracy and low loss values obtained during both training and validation indicate that the model learned temporal gesture patterns effectively without exhibiting substantial performance degradation on unseen validation data. This result suggests that the LSTM architecture achieved stable learning behavior and good generalization capability during the training process.

To visualize the learning behavior of the model during training, the accuracy and loss progression across training epochs are presented in Figure 3.

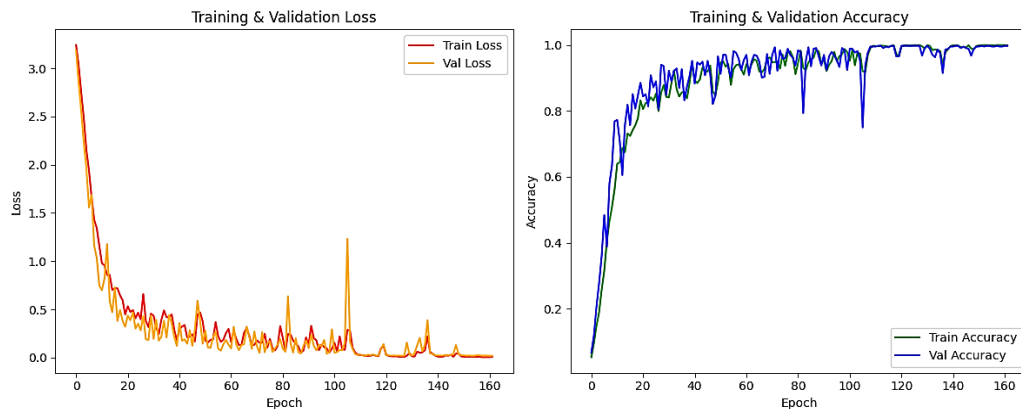


Figure 3. Training and Validation Accuracy-Loss Curves

As shown in Figure 3, the training curves demonstrate stable convergence throughout the optimization process. Both training and validation accuracy increased consistently, while the corresponding loss values decreased gradually as the number of epochs increased. The absence of significant fluctuations between training and validation performance indicates that the model was able to learn temporal gesture representations effectively under the given training configuration.

To further validate the reliability of the training results, 10-fold cross-validation was conducted during model development. The performance obtained from each validation fold is summarized in Table 3.

Table 3. Validation Performance Across 10-Fold Cross-Validation

Fold	Accuracy (%)	Loss
1	100	0.0051
2	100	0.0064
3	100	0.0048
4	100	0.0048
5	99.54	0.0167
6	99.54	0.0167
7	99.08	0.0541
8	99.08	0.0715
9	99.77	0.0189
10	99.77	0.0225
Avg	99.68	0.0221

The values presented in Table 3 correspond to the best validation results obtained from each fold during the cross-validation process. The model achieved consistently high validation performance across all folds, with accuracy values ranging from 99.08% to 100% and an average accuracy of 99.68%. The relatively low loss values observed in each fold indicate stable optimization and effective learning behavior throughout the training process.

Although slight performance variations were observed among several folds, the overall results remained highly consistent across different data partitions. These findings show that the developed LSTM architecture was able to learn temporal gesture patterns reliably and that the obtained validation performance was not dependent on a specific train-validation split configuration.

3.2. Inference Performance

The trained model was evaluated using unseen gesture sequences to measure its recognition capability under inference conditions. Performance assessment was conducted using several classification evaluation metrics with macro averaging to ensure balanced contribution from all alphabet classes. The quantitative results obtained during the inference stage are summarized in Table 4.

Table 4. Inference Performance of the Proposed Model

Metric	Results (%)
Accuracy	89.42
Precision	91.57
Recall	89.42
F1-Score	88.74

As shown in Table 4, the model obtained an inference accuracy of 89.42% on unseen test data. The precision value reached 91.57%, while recall was recorded at 89.42%, resulting in an overall F1-score of 88.74%. The relatively balanced values of precision, recall, and F1-score indicate that the proposed model maintained consistent classification performance across different alphabet classes. The high precision suggests that most predicted gestures were correctly classified, while the comparable recall value indicates that the majority of actual gesture classes were successfully detected. These results demonstrate that the model achieved reliable recognition performance without favoring specific gesture classes.

To evaluate computational efficiency and real-time capability, inference time analysis was conducted using 312 evaluation samples. The inference time evaluation results are presented in Table 5.

Table 5. Inference Time Evaluation Results

Parameter	Value
Total Inference Time	40.3901 seconds
Average FPS	249.18 frames/second
Average Latency	129.46 ms/sample

Based on Table 5, the framework achieved an average processing speed of 249.18 FPS with an average latency of 129.46 ms per sample. These results indicate that the model is capable of performing rapid gesture sequence inference with low processing delay, supporting real-time implementation requirements. The high FPS performance demonstrates that the LSTM-based temporal modeling framework can efficiently process sequential hand keypoint data while maintaining stable recognition performance on unseen gesture sequences.

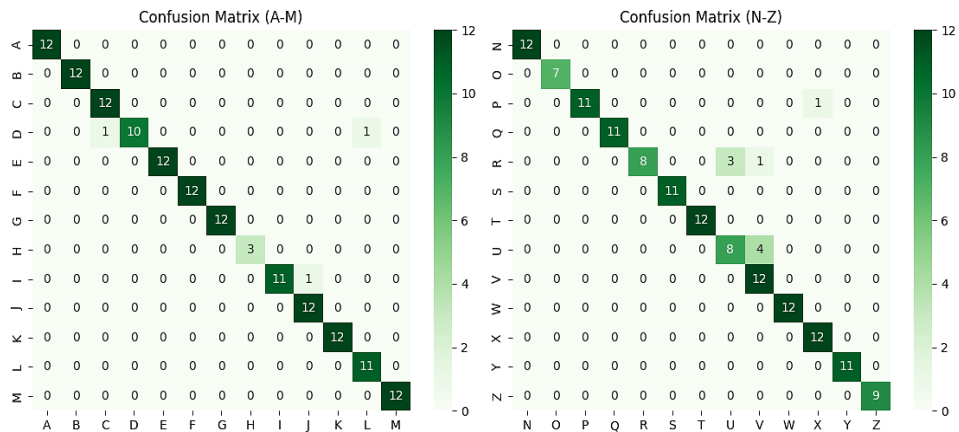


Figure 4. Confusion Matrix of Inference Testing Results

To further examine class-level recognition performance, a confusion matrix was generated to visualize the distribution of predicted and actual gesture classes. Figure 4 presents the confusion matrix obtained from the inference testing process.

As shown in Figure 4, most alphabet classes were classified correctly, as indicated by the strong concentration of predictions along the main diagonal. Misclassifications were primarily observed among gestures with similar hand configurations or motion characteristics, suggesting that visual and temporal similarity remains a major challenge for certain alphabet pairs. Nevertheless, the overall distribution demonstrates that the proposed model was able to distinguish most gesture classes effectively.

3.3. Statistical Validation Using ANOVA

To analyze system robustness under different environmental conditions, repeated-measures ANOVA was conducted to analyze the effects of user variation, camera distance, and lighting intensity on system performance. The statistical summary is presented in Table 6.

Table 6. Summary of Repeated-measures ANOVA Results

Factor	F-value	p-value	Effect Size (η^2)
User	1.19	0.39	0.0052 (0.52%)
Camera Distance	100.60	< 0.01	0.96 (96%)
Lighting Intensity	1.81×10^5	0.015	0.0095 (0.95%)
Distance $\times$ Lighting	1.35×10^2	0.0073	0.0159 (1.59%)

As shown in Table 6, the repeated-measures ANOVA indicates variability in statistical significance and effect magnitude across the evaluated factors. Among the examined variables, camera distance exhibits the largest effect size, indicating the strongest statistical influence on system performance within the tested conditions. The extremely large effect size observed for camera distance ($\eta^2 = 0.96$) suggests that recognition performance is highly dependent on the spatial resolution of captured hand landmarks. As the distance between the user and the camera increases, hand regions occupy fewer pixels, potentially reducing landmark localization accuracy and affecting gesture classification performance.

The reported p-values reflect hypothesis testing outcomes under the predefined significance threshold ($\alpha = 0.05$), whereas the η^2 values quantify the proportion of variance in performance metrics attributable to each factor. The inclusion of both main effects and interaction effects provides a structured statistical overview of performance variability across experimental conditions within the within-subject design.

3.4. Contextual Word Prediction Performance

Contextual word prediction performance was evaluated by comparing two configurations: (1) LSTM combined with Trie-based lexical constraints and (2) LSTM combined with Trie and IndoBERT. Table 7 presents the comparison results.

Table 7. Contextual Word Prediction Performance

Model Configuration	Accuracy (%)
LSTM + Trie	18.2
LSTM + Trie + IndoBERT	58.64
Improvement	$\uparrow 3.2\times$

As shown in Table 7, the integration of IndoBERT substantially improved contextual word prediction accuracy from 18.2% to 58.64%, representing a 3.2-fold increase over the Trie-only configuration. This improvement indicates that lexical filtering alone is insufficient to resolve ambiguity during partial word formation, particularly when multiple candidate words share similar prefixes. By incorporating contextual information from the surrounding sentence, IndoBERT is able to prioritize semantically relevant candidates, resulting in more accurate word prediction outcomes.

3.5. McNemar Test Analysis

To statistically assess the impact of contextual language modeling, a McNemar test was performed using paired prediction outcomes obtained from identical contextual word prediction scenarios. The analysis involved 220 paired observations generated from two target words ("Langit" and "Teman"), where each spelling stage was evaluated through 20 repeated trials. For every observation, prediction outcomes from the LSTM + Trie and LSTM + Trie + IndoBERT configurations were recorded as either correct or incorrect, allowing direct pairwise comparison of the two approaches. Table 8 shows the McNemar test results.

Table 8. McNemar Test Results

Comparison	χ^2	p-value
LSTM + Trie vs LSTM + Trie + IndoBERT	87.01	7.27×10^{-21}

As shown in Table 8, the McNemar test demonstrates a statistically significant difference between the two contextual prediction approaches under the predefined significance threshold ($\alpha = 0.05$). The test was conducted using paired classification outcomes to evaluate whether the observed performance variation between configurations reflects a consistent difference in prediction behavior.

The statistical evidence confirms that the performance improvement obtained after integrating IndoBERT was consistently observed across paired prediction samples and was unlikely to be caused by random variation. This finding strengthens the conclusion that contextual language modeling contributes meaningfully to word prediction performance and provides quantitative support for the effectiveness of IndoBERT within the framework.

3.6. Real-Time System Performance

Real-time system testing was conducted using webcam-based input to evaluate the operational performance of the framework. Gesture recognition was performed after collecting 30 consecutive frames for each gesture sequence. The system was able to maintain responsive gesture recognition performance with an average processing latency below one second during real-time operation.

For contextual word prediction, additional latency was observed during the early spelling stages, particularly when only a few letters had been detected. In this condition, the Trie structure generated a large number of candidate words that needed to be re-ranked using IndoBERT, resulting in prediction latency of approximately 2–3 seconds. However, as more letters were accumulated, the number of candidate words decreased significantly, allowing the contextual prediction process to operate faster with latency reduced to below one second. Figure 5 illustrates the real-time user interface during gesture detection and prediction.

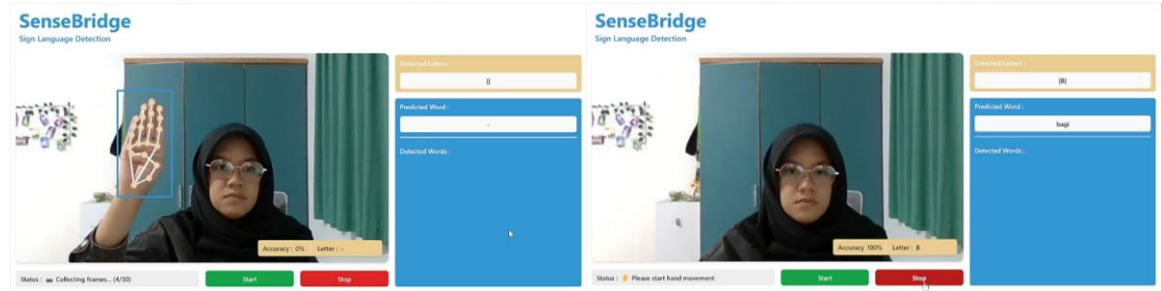


Figure 5. Real-Time Operation Interface of the Sign Language Recognition System.

Overall, the system maintained stable real-time gesture recognition performance, while the contextual prediction module exhibited adaptive latency depending on the number of candidate words generated during early spelling stages.

3.7. Discussions

The experimental findings demonstrate that the proposed framework successfully integrates temporal gesture modeling and contextual language modeling to address key limitations in existing SIBI recognition systems. The LSTM-based model achieved a training accuracy of 99.71% and a validation accuracy of 99.77%. In addition, the 10-fold cross-validation experiment yielded an average validation accuracy of 99.68% with consistently low loss values across all folds. These findings indicate that the architecture learned stable temporal representations and maintained consistent performance across different data partitions, reducing the likelihood of overfitting or bias from a particular data split. When evaluated on unseen users, the model achieved an inference accuracy of 89.42%. The observed performance decrease of approximately 10% is expected in real-world sign language recognition scenarios, where variations in hand shape, signing style, motion speed, and environmental conditions are more pronounced than in the training data. Therefore, the reduction in performance is more likely attributable to increased user and environmental variability rather than overfitting, while the achieved accuracy still indicates satisfactory generalization capability for real-time SIBI recognition.

This performance difference can be attributed primarily to environmental variability during testing rather than model instability. While training data were collected under consistent camera positioning and controlled hand-to-camera distance, the testing phase involved a broader range of users and more varied spatial configurations. The repeated-measures ANOVA results confirm that camera distance has a strong and statistically significant effect on recognition performance ($F = 100.6$, $p < 0.01$, $\eta^2 = 0.96$), representing a large

effect size. Increased hand-to-camera distance reduces spatial resolution and landmark clarity, leading to less precise keypoint extraction and consequently affecting classification accuracy. In contrast, user variation did not produce a significant effect ($F = 1.19$, $p = 0.39$, $\eta^2 = 0.0052$), indicating that the model generalizes well across different individuals. Therefore, the observed training–testing gap reflects environmental sensitivity rather than overfitting to specific users.

The proposed framework also demonstrated high computational efficiency during real-time evaluation. Gesture sequences were processed at an average speed of 249.18 FPS with an average latency of 129.46 ms per sample. The obtained results show that the keypoint-based LSTM architecture can perform rapid gesture recognition with minimal processing delay. Furthermore, the combination of lightweight landmark-based features and temporal sequence modeling allows the system to maintain real-time operation without requiring high-end computational resources.

A key contribution of this study lies in contextual language integration. The Trie-only configuration achieved 18.2% accuracy, highlighting the limitations of prefix-based lexical filtering without semantic modeling. After incorporating IndoBERT, word-level accuracy increased to 58.64%. Although the absolute accuracy remains moderate, the relative improvement of more than three times demonstrates the substantial impact of contextual semantic modeling. The McNemar test ($\chi^2 = 87.01$, $p < 0.001$) further confirms that this improvement is statistically significant and not attributable to random variation.

The moderate absolute contextual accuracy can be explained by the complexity of word reconstruction from alphabet-level predictions. Small misclassifications at the gesture level can propagate and alter the final word prediction, particularly during early spelling stages when contextual information is limited. As more letters accumulate, contextual probability distribution becomes more stable, which explains the observed reduction in prediction latency and improved selection confidence during later stages of spelling. Therefore, the contextual module should be interpreted as a semantic refinement mechanism rather than a standalone correction system.

From a real-time usability perspective, the system demonstrates practical responsiveness. Gesture recognition is completed in under one second after collecting 30 frames, while contextual word prediction exhibits adaptive latency depending on the number of detected letters. During early spelling stages, IndoBERT introduces additional computational overhead due to the large number of candidate words generated by the Trie structure. However, as more letters are accumulated, the candidate space becomes smaller, allowing prediction latency to decrease to near real-time performance, often below one second. Despite the additional processing cost, IndoBERT significantly improves semantic coherence and candidate selection stability.

Nevertheless, environmental sensitivity remains a notable limitation. The strong dependence on camera distance suggests that performance may degrade in unconstrained scenarios with low-resolution input or significant spatial variation. Future work may address this limitation through adaptive spatial normalization, distance-aware scaling mechanisms, or the integration of lightweight transformer-based contextual models to reduce computational latency. Additionally, expanding the dataset to include greater environmental diversity and transitioning from alphabet-level modeling to continuous sentence-level recognition represent important next steps for improving robustness and practical applicability.

Overall, the results indicate that integrating LSTM-based temporal modeling with contextual language modeling significantly enhances both gesture recognition robustness and semantic interpretability. While challenges related to environmental variability and contextual complexity remain, the framework represents a substantial advancement beyond isolated alphabet classification toward meaningful, real-time SIBI communication systems for inclusive assistive technology deployment.

4. CONCLUSION

This study successfully developed a real-time Indonesian Sign Language (SIBI) recognition system by integrating LSTM-based temporal modeling and IndoBERT contextual language modeling. The LSTM framework achieved 89.42% accuracy, 91.57% precision, 89.42% recall, and 88.74% F1-score while maintaining high computational efficiency with an average processing speed of 249.18 FPS and an average latency of 129.46 ms per sample. The results demonstrate that keypoint-based temporal modeling can effectively recognize dynamic sign language gestures with good computational efficiency. Error analysis showed that most recognition errors occurred in gesture classes with similar hand configurations and temporal movement patterns. In addition, recognition performance decreased under larger hand-to-camera distances, indicating that landmark quality and spatial resolution remain important factors affecting classification accuracy.

The integration of IndoBERT significantly improved contextual word prediction performance compared to lexical filtering alone, indicating that contextual language understanding contributes substantially to semantic interpretation in sign language translation. Furthermore, statistical evaluation confirmed that camera distance has a significant influence on recognition performance, while user variation showed relatively small impact, demonstrating acceptable model generalization across different users.

The proposed framework has potential applications in assistive communication systems, inclusive educational technology, and real-time accessibility tools for deaf individuals. The integration of temporal gesture recognition and contextual semantic modeling provides a more meaningful communication mechanism compared to isolated gesture classification systems. Future research may focus on improving robustness under unconstrained environmental conditions, reducing contextual prediction latency, and expanding the dataset with more diverse gesture variations. Further development toward sentence-level recognition and continuous sign language translation may also improve practical usability in real-world communication scenarios.

ACKNOWLEDGMENTS

The authors would like to express their sincere appreciation to the teachers and students of the Special Education School (SLB) for their valuable participation and support during the data collection process.

REFERENCES

- [1] A. Podury, N. T. Jiam, M. Kim, J. I. Donnenfield, and A. Dhand, "Hearing and sociality: the implications of hearing loss on social life," *Front. Neurosci.*, vol. 17, no. 1245434, Oct. 2023, doi: 10.3389/fnins.2023.1245434.
- [2] P. A. Rodríguez-Correa, A. Valencia-Arias, O. N. Patiño-Toro, Y. Oblitas Díaz, and R. Teodori De la Puente, "Benefits and development of assistive technologies for Deaf people's communication: A systematic review," *Front. Educ. (Lausanne)*, vol. 8, no. 1121597, Apr. 2023, doi: 10.3389/educ.2023.1121597.
- [3] S. Syamsuddin, T. Pristiwaluyo, W. A. Saleh, and Z. Zulfitriah, "Development of Indonesian Sign Language System (SIBI) Dictionary Application for Students with Special Needs," *AL-ISHLAH: Jurnal Pendidikan*, vol. 15, no. 3, pp. 3132–3143, Sep. 2023, doi: 10.35445/alishlah.v15i3.3529.
- [4] K. M. Concha, F. P. Zenteno, and J. T. Alfaro, "Use of artificial intelligence in libraries a systematic review, 2019-2023," *South African Journal of Library and Information Science*, vol. 90, no. 2, Aug. 2024, doi: 10.7553/90-2-2387.
- [5] D. Vaněček, Y. I. Yorulmaz, and D. Dobrovská, "A holistic exploration of student attitudes toward AI use in higher education: an international comparison," *Cogent Education*, vol. 12, no. 1, 2025, doi: 10.1080/2331186X.2025.2571691.
- [6] Asma Syafira and R. A. Yuana, "Interactive Mobile Applications for Social Communication in Deaf and Hearing-Impaired Populations: A Systematic Review," *International Journal of Interactive Mobile Technologies (IJIM)*, vol. 20, no. 02, pp. 78–95, Jan. 2026, doi: 10.3991/ijim.v20i02.57207.
- [7] M. Sari and E. R. Jamzuri, "Hand Sign Recognition of Indonesian Sign Language System (SIBI) Using Inception V3 Image Embedding and Random Forest," *Jurnal Rekayasa Sistem dan Teknologi Informasi*, vol. 9, no. 2, pp. 258–265, Apr. 2025, doi: 10.29207/resti.v9i2.6156.
- [8] S. Shivani and S. Bal Gupta, "A Comprehensive Analysis of Recognition of Hand Gestures using Machine Learning," *Makara Journal of Technology*, vol. 29, no. 1, pp. 48–60, Apr. 2025, doi: 10.7454/mst.v29i1.1679.
- [9] J. Huang and V. Chouvatut, "Video-Based Sign Language Recognition via ResNet and LSTM Network," *J. Imaging*, vol. 10, no. 6, p. 149, Jun. 2024, doi: 10.3390/jimaging10060149.
- [10] K. Xia, W. Lu, H. Fan, and Q. Zhao, "A Sign Language Recognition System Applied to Deaf-Mute Medical Consultation," *Sensors*, vol. 22, no. 23, p. 9107, Dec. 2022, doi: 10.3390/s22239107.
- [11] K. Navendu and V. Sahula, "Word Level Sign Language Recognition using MediaPipe and LSTM-GRU Network," in *2024 IEEE International Symposium on Smart Electronic Systems (iSES)*, IEEE, Jul. 2024, pp. 13–18. doi: 10.1109/iSES63344.2024.00014.
- [12] T. Purbojo and A. P. Wijaya, "Enhancing Pose-Based Sign Language Recognition: A Comparative Study of Preprocessing Strategies with GRU and LSTM," *Advance Sustainable Science, Engineering and Technology*, vol. 7, no. 2, Feb. 2025, doi: 10.26877/asset.v7i2.1658.
- [13] I. V. Lemmuela, M. Ayub, and O. Karnalim, "Dynamic Sign Language Recognition in Bahasa using MediaPipe, Long Short-Term Memory, and Convolutional Neural Network," *Journal of Information Systems Engineering and Business Intelligence*, vol. 11, no. 1, pp. 17–29, Feb. 2025, doi: 10.20473/jisebi.11.1.17-29.
- [14] E. Altirika and W. P. Sari, "Pengembangan Deteksi Realtime untuk Bahasa Isyarat Indonesia dengan Menggunakan Metode Deep Learning Long Short Term Memory dan Convolutional Neural Network," *Jurnal Teknologi Informatika dan Komputer*, vol. 9, no. 1, pp. 1–13, Mar. 2023, doi: 10.37012/jtik.v9i1.1272.
- [15] P. Ho and H. Santoso, "LSTM-Based Hand Gesture Recognition for Indonesian Sign Language System (SIBI) on Affix, Alphabet, Number, and Word," *Journal of Applied Informatics and Computing (JAIC)*, vol. 9, no. 3, p. 928, Jun. 2025, doi: 10.30871/jaic.v9i3.9607.

-
- [16] C. A. Sari, E. H. Rachmawanto, Z. Saifullah, C. Jatmoko, and D. Sinaga, "Real-time detection of indonesian sign language (ISL) gestures based on long short-term memory," *Journal of Soft Computing Exploration*, vol. 5, no. 3, pp. 251–262, Sep. 2024, doi: 10.52465/josce.v5i3.452.
 - [17] D. Bala, B. Sarkar, M. I. Abdullah, M. A. Hossain, and I. Abdullah, "American Sign Language Alphabets Recognition using Convolutional Neural Network," *Article in International Journal of Knowledge Based Computer Systems*, vol. 6, no. 1, pp. 33–40, Jun. 2021, [Online]. Available: <https://www.researchgate.net/publication/352878275>
 - [18] G. H. Samaan *et al.*, "MediaPipe's Landmarks with RNN for Dynamic Sign Language Recognition," *Electronics (Switzerland)*, vol. 11, no. 19, p. 3228, Oct. 2022, doi: 10.3390/electronics11193228.
 - [19] W. Jintanachaiwat *et al.*, "Using LSTM to translate Thai sign language to text in real time," *Discover Artificial Intelligence*, vol. 4, no. 1, Dec. 2024, doi: 10.1007/s44163-024-00113-8.
 - [20] R. R. Agustin, H. Maulana, and E. P. Mandyartha, "Detection of Actions BISINDO (Indonesian Sign Language) Into Text-to-Speech using Long Short-Term Memory with Mediapipe Holistics," *Jurnal Teknik Informatika (Jutif)*, vol. 5, no. 4, pp. 1051–1061, Nov. 2023, doi: 10.52436/1.jutif.2024.5.4.1492.
 - [21] L. Katriel and R. Wiguna, "Sentence-Level Indonesian Sign Language (BISINDO) Recognition Using 3D CNN-LSTM and 3D CNN-BiLSTM Models (Indonesian Sign Language); 3D Convolutional Neural Network (3D CNN); Long Short-Term Memory (LSTM) Network; Bidirectional Long Short-Term Memory (BiLSTM); Connectionist Temporal Classification (CTC)," *IJACSA International Journal of Advanced Computer Science and Applications*, vol. 16, no. 9, pp. 433–443, 2025, doi: 10.14569/IJACSA.2025.0160941.
 - [22] Y. Setiawan and L. A. Wulandhari, "Comparative Analysis of IndoBERT and LSTM for Multi-Label Text Classification of Indonesian Motivation Letter," *JOIN (Jurnal Online Informatika)*, vol. 10, no. 27, pp. 260–269, 2025, doi: 10.15575/join.v10i2.1499.
 - [23] D. Setiawan, D. U. Iswavigra, and E. Anggiratih, "Implementation of IndoBERT for Sentiment Analysis of the Constitutional Court's Decision Regarding the Minimum Age of Vice Presidential Candidates," *Scientific Journal of Informatics*, vol. 12, no. 3, pp. 397–406, 2025, doi: 10.15294/sji.v12i3.26360.
-