LSTM-Based Hand Gesture Recognition for Indonesian Sign Language System (SIBI) on Affix, Alphabet, Number, and Word

Patricia Ho[1]*, Handri Santoso[2]*
*Informatics, Science and Technology Faculty, Pradita University
patricia.ho@student.pradita.ac.id[1], handri.santoso@pradita.ac.id[2]

Article Info

Article history:

Received 2025-05-05
Revised 2025-06-14
Accepted 2025-06-17

Keyword:

Deep Learning,
Hand Gesture Recognition,
Indonesian Sign Language System (SIBI),
LSTM.

ABSTRACT

Sign language plays a critical role in enabling communication for the Deaf and hard-of-hearing community in Indonesia, yet there remains a significant gap in technological support for recognizing the official Indonesian sign language, Sistem Isyarat Bahasa Indonesia (SIBI). This study presents a deep learning-based hand gesture recognition system for SIBI, focusing on four primary gesture categories: affix, alphabet, number, and word. A large and diverse dataset of 21,351 videos was collected, covering 18 affix, 26 alphabet, 35 number, and 29 word classes. Hand keypoints were extracted using MediaPipe Holistic, and a bidirectional long short-term memory (BiLSTM) model was trained using 5-fold stratified cross-validation. The model achieved high recognition performance in the alphabet, number, and word categories, with mean test accuracies of 93.94%, 91.48%, and 92.41%, respectively, and slightly lower performance in the affix category at 68.17%. The affix category posed particular challenges due to subtle hand shape differences and high variability between signers, while the alphabet category consistently showed the highest accuracy due to its distinct and static handshapes. Evaluation metrics, including precision, recall, F1-score, and confusion matrix analysis, provided further insights into model strengths and limitations. Overall, the study demonstrates the effectiveness of LSTM models for sequential hand gesture recognition in SIBI and highlights areas for future improvement, such as handling non-manual features and improving generalization across signers.

I. Introduction

According to Indonesia's 2022 Population Census (Badan Pusat Statistik, BPS), out of 253.7 million people aged five and above, over 255,000 are completely unable to hear, while around 669,000 experience severe hearing difficulties, and over 4 million report mild hearing difficulties. Together, it shows that nearly 1 million Indonesians live with severe or total hearing loss, forming a substantial deaf and hard-of-hearing community [1].

Recent research shows that despite Indonesia's progressive disability policies, such as Law No. 8 of 2016, the country remains far from inclusive at the community level. Public understanding of the needs of disabled groups, especially the Deaf who rely on sign language, is still low, often leading to unrealistic expectations like forcing oral communication. The gap in infrastructure between big cities like Surabaya and regions such as Lampung continues to limit access and worsen social isolation. To develop a more inclusive society, Indonesia will need to not only improve the infrastructure across the country, but also include sign language education and disability awareness into schools. This will grow a generation who understands, respects and advocates for the rights of people with disabilities [2].

Sign language is a completely developed visual language that uses movement of the hands, facial expressions and body language to express meaning. For the Deaf community, it is not only a communication tool, but also part of their cultural and language identity [3]. In Indonesia, the official sign language is called Sistem Isyarat Bahasa Indonesia (SIBI), which is different than traditional sign languages like BISINDO in that it includes affixes, alphabets, numbers, and everyday vocabulary. SIBI is commonly used in both classroom and special needs education settings, and has also been adopted as a possible means of written and verbal communication on television [4]. SIBI can seem a bit rigid in informal conversation, but since it uses the same format as Indonesian grammar, it also generally does further literacy [5].

Despite the importance of sign language, communication gaps remain high. For example, in a study conducted in 2019 in a special needs school in Pekanbaru, it was found that Deaf parents were having communication difficulties with levels ranging from moderate to severe, mostly because of their ability to sign, and their inability to express messages, or receive messages. The communication difficulties raised parental stress levels, while the analysis reported a strong positive relationship between communication difficulties and stress at a correlation value of 0.819, indicating important communication tools and access to sign language in parts of Indonesia, are required [6].

The application of technology in the form of deep learning and computer vision represents an alternative way for reducing communication barrier in Deaf and hearing people. In previous research, convolutional neural networks were developed using a dataset of 24 alphabetic gestures, excluding J and Z as they are difficult to represent because of their gestures. The models performed well with a training accuracy of 99.7%, a test accuracy of 87.5% for forty centimeters, and 79.2% for sixty centimeters. The models used MediaPipe to detect hand landmarks, but their usefulness relies on their operation being uninhibited, when working on longer distances, and more nuanced gestures [7].

Previous studies have also applied deep learning models for SIBI recognition, including the use of LSTM for inflectional words such as affixes [8], hybrid CNN-LSTM architectures for real-time SIBI recognition [9], and LSTM combined with MediaPipe for symbol detection [10].

In an effort to advance the effectiveness of sign language translation through SIBI, recent models have emerged that combine the manual features, including the use of hand gestures, and the non-manual features, including facial expressions. By recognizing visual and emotional cues, these studies put forth a multi-level model consisting of detection, segmentation, recognition, and translation and finally reaching spoken language outputs that are more accurate and naturalistic [11]. Since SIBI uses root words and affixes that adhere to Indonesian grammatical rules, segmenting continuous sign language sequences is a significant challenge that these studies attempt to address. In order to address this, a temporal action segmentation technique based on optical flow was created. It uses the Farneback algorithm to detect motion boundaries between signs without requiring the high processing overhead of deep learning models. Strong segmentation performance was shown by experimental findings using dense optical flow on SIBI videos, with optimal Perf and F1r scores of 0.8298 and 0.8524, respectively [12].

Although recent research has combined facial expressions and hand gestures to enhance SIBI-based sign language translation, this study only looks at hand gesture recognition. The goal of the research is to improve SIBI identification systems' accuracy and dependability by focusing solely on manual components rather than non-manual ones. Affixes, Alphabet, Number, and Word are the four primary SIBI categories that are especially examined in this study. In the context of SIBI, affix gestures carry unique linguistic and visual complexity. Unlike the Alphabet or Number categories that often have distinct and easily separable gestures, affix gestures such as "se," "me," and "ter" involve only subtle differences in hand shape or thumb placement, despite having similar hand orientation and movement. For example, "se" and "me" both use two hands, but the thumb is placed inside the fist for "me" and outside for "se." Similarly, "ter" and "me" share the same direction and general form, but differ slightly in the placement of the thumb beneath the index finger in "ter," where the index finger bends over the thumb. These fine-grained differences make affix gestures particularly challenging to distinguish, even for human observers, and therefore present a major challenge for automated recognition systems. This study refers to the Kamus SIBI published by the Indonesian Ministry of Education and Culture to understand and incorporate these gesture definitions and nuances into the research [13].

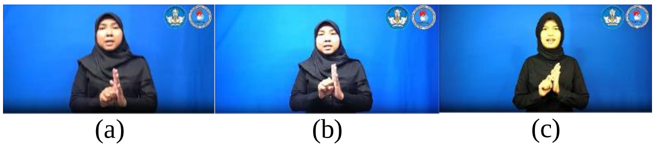


Figure 1. Visual comparison of the affix gestures "me" (a), "se" (b), and "ter" (c) in SIBI, focusing on hand shape differences.

While these studies demonstrate the potential of deep learning in SIBI recognition, they typically focus on specific components such as affixes or a limited set of gestures and do not systematically compare performance across categories. This study addresses that gap by evaluating LSTM performance across the four primary SIBI categories, which are Affix, Alphabet, Number, and Word, using different dataset that is larger and more diverse in classes.

II. Method

This study began by identifying the main issue, which is communication challenges faced by the Deaf community in Indonesia, especially around accessibility and sign language understanding. Interviews with three Deaf or Hard of Hearing individuals and representatives from the AUDISI Foundation provided insights into the social and professional barriers they face and their positive view of technology as a tool for inclusion. The research then moved through stages of data acquisition, preprocessing, model training, and evaluation, as outlined in Figure 2.

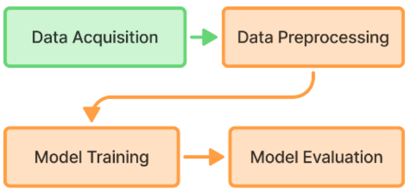


Figure 2. General Research Implementation Flow

A. *Data Acquisition*

The dataset used in this study consists of Indonesian Sign Language System (SIBI) videos compiled from the official online dictionary provided by the Ministry of Education and Culture of the Republic of Indonesia. This dataset, used with permission from the data owner, was carefully selected to support research on hand gesture action recognition [14]. It includes four main SIBI categories, which are affixes (18 classes), alphabets (26 classes), numbers (35 classes), and words (29 classes). Each class represents a distinct sign form and has been organized into individual word units.

The dataset comprises a total of 21,351 video files. Data collection involved 22 subjects for the sentence category and 20 subjects for the affix, alphabet, and number categories, consisting of a mix of teachers and students. For the sentence data, only videos with neutral facial expressions were included, focusing on ten sentences commonly used in conversational contexts, which were segmented into word-level units.

Additionally, sequence lengths before normalization and after cleaning the intro and outro segments were analyzed, with detailed frame counts available in the attached metadata files for each category (affix, alphabet, number, word). This preprocessing step ensures consistency in temporal features across samples and improves the reliability of spatial-temporal feature learning.

TABLE I
DATASET DISTRIBUTION OF THE DATASET USED IN RESEARCH

Category	Number of Classes	Number of Videos per Class	Total Data
Affix	18	200	3600
Alphabet	26	185-200	5167
Number	35	179-200	6785
Word	29	199-200	5799

B. *Data Acquisition*

The process begins with the existing SIBI dataset, which includes affix, alphabet, number, and sentence categories. To align the data with the study's focus on word-level hand gesture recognition, the sentence videos were manually cropped into individual word units. The preprocessing stage then involved two main steps. First, the video data was cleaned by removing unnecessary intro and outro sections, resizing frames to 224 by 224 pixels, converting frames to grayscale and applying CLAHE, sharpening, Gaussian blurring, and then converting them back to RGB. Hand keypoints were extracted using MediaPipe Holistic, which offers more robust and stable detection by incorporating full-body context compared to using the hand module alone, improving accuracy in complex sign gestures [15], [16].

Before normalization, the cleaned dataset included 21,351 videos across the four SIBI categories, with substantial variability in sequence lengths, as summarized in Table X. This variation underscores the need for robust temporal normalization to standardize input length across samples and improve the consistency of spatial-temporal feature learning.

TABLE II
SUMMARY OF FRAME STATISTICS BEFORE NORMALIZATION

Category	Min Frames	Max Frames	Mean Frames	Std Frames
Affix	10	57	29.7	8.3
Alphabet	5	72	5167	9.5
Number	6	69	6785	9.8
Word	3	56	23.8	7.7

In the next step, the dataset was refined by normalizing the frame sequences to 30 frames per video, re-extracting and interpolating keypoints to address missing data, and applying a binary mask to finalize the input format. The masking process was validated through visualization of keypoint overlays before and after masking. The blue ticks in the visualization represent right-hand keypoints, which were still included as valid in several samples due to imperfect masking. This inclusion may introduce irrelevant movement information from the right hand during training, potentially adding noise and reducing classification accuracy, especially in signs performed primarily with the left hand.

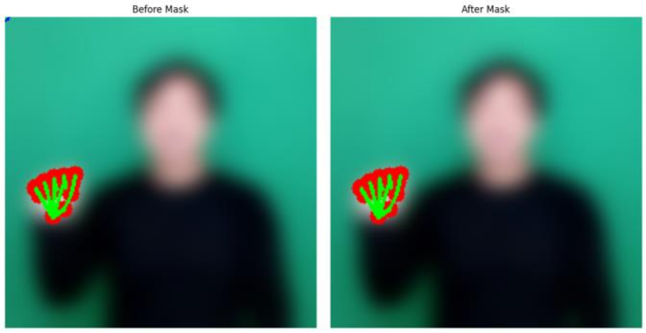


Figure 3. Comparison of hand keypoint visualization before and after applying the masking process.

This preprocessing pipeline resulted in a cleaned SIBI dataset structured into four key categories, consisting of 18 affix classes, 26 alphabet classes, 35 number classes, and 29 word classes, all prepared for the subsequent model training and evaluation phases.

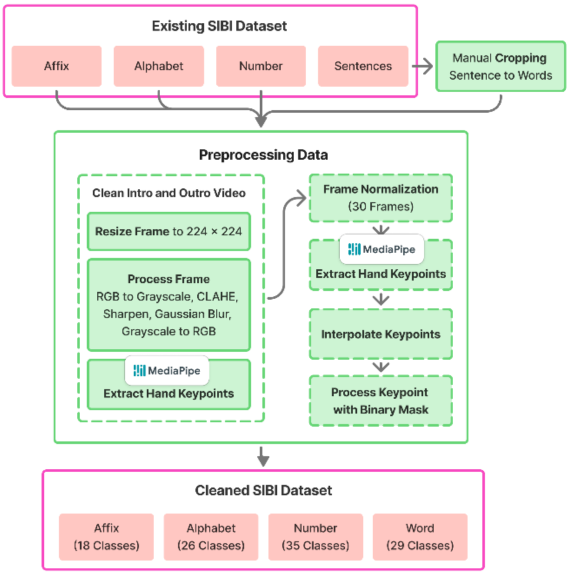


Figure 4. Data Preprocessing Flow

C. *Model Training*

The overall model training procedure used in this study is illustrated in Figure 5. The model receives two primary inputs: the hand keypoint sequences and their corresponding binary masks. Each input is shaped as a batch with two coordinate values (x and y) obtained from MediaPipe, two hands, 21 keypoints per hand, and a sequence of frames [17]. The first processing step multiplies the keypoint input by the unsqueezed binary mask, effectively zeroing out missing or invalid keypoints. The (2 hands × 21 keypoints × 2 coordinates) structure is then flattened into a per-frame vector of 84 features.

The sequential input is processed by a bidirectional long short-term memory (BiLSTM) network composed of two stacked layers, each with a hidden size of 96 units, resulting in a 192-dimensional output vector per time step. The BiLSTM extends the traditional LSTM by processing the input sequence in both forward and backward directions [18], enabling the model to learn dependencies between past and future frames. This bidirectional mechanism enhances gesture recognition by providing context from the entire sequence, improving classification accuracy.

The temporal output is then passed to a fully connected classification layer consisting of a linear transformation from 192 to 256 dimensions, followed by a ReLU activation, a dropout layer with a rate of 0.3 to prevent overfitting [19], [20], and a final linear layer projecting to the number of gesture categories. During training, the output logits are used to compute the cross-entropy loss in combination with a softmax activation, guiding the optimization of model parameters through backpropagation[21].

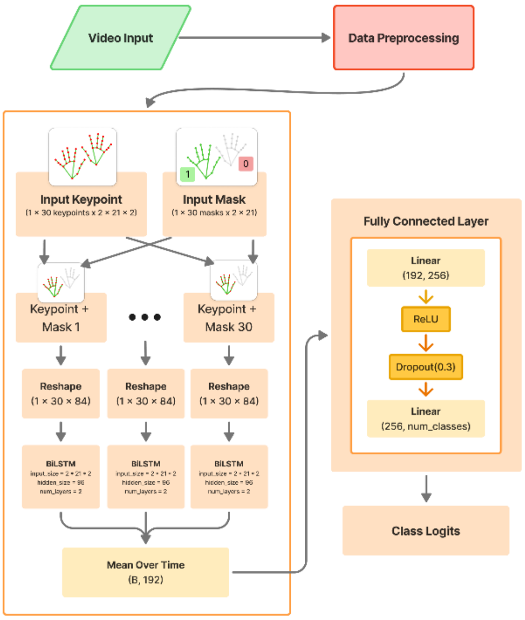


Figure 5. Model Training Flow

The model optimization was performed using the Adam optimizer with an initial learning rate of 1×10^{-4}. To improve convergence, a step learning rate scheduler was applied, reducing the learning rate by a factor of 0.5 every 5 epochs. Training was conducted over 100 epochs per fold with a batch size of 16, using 5-fold stratified cross-validation to ensure robust generalization across data splits. All training procedures were executed using GPU acceleration on an NVIDIA GeForce RTX 2070 with Max-Q Design.

D. *Model Evaluation*

The performance of the model is evaluated using different classification metrics commonly utilized in multi-class contexts, such as but not limited to accuracy, loss, precision, recall, F1-score, and the confusion matrix.

Accuracy is defined as the fraction of correct predictions compared to the total amount of data [22].

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Predictions} \tag{1}$$

Additionally, precision and recall are used to measure how well the model predicts each class. Precision refers to the proportion of correct positive predictions out of all predictions made for a class, while recall refers to the proportion of correct positive predictions out of all actual instances of that class [23]. he formulas for these metrics use TP (true positives), FP (false positives), and FN (false negatives).

$$Precision = \frac{TP}{TP+FP} \tag{2}$$

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

The F1-score is a combined metric that balances precision and recall by taking the harmonic mean of the two.

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

The confusion matrix is a useful evaluation tool that shows both correct and incorrect predictions across all classes. In this matrix, the rows represent the predicted classes, while the columns show the actual classes. It gives more detailed insights than just overall accuracy by breaking down true positives, false positives, and false negatives. In multiclass tasks, the confusion matrix plays an important role in providing a fair and balanced view of how well the system can tell different gesture categories apart [24].

In this study, evaluation is performed on a per gesture sequence basis, meaning that the model predicts and is assessed on the entire sequence of frames representing a gesture or word, not on individual frames or isolated keypoints. This approach ensures the temporal dynamics of the gestures are taken into account during evaluation.

III. Result and Discussion

A. Data Preparation and Splitting

The data set was created with hand gesture video recordings that are adherent to the Indonesian Sign Language System or SIBI. Every sample consists of 30 frames, with hand keypoints structured as 2 hands, 21 points for each hand, and x and y coordinates. There is the use of a binary mask to indicate valid keypoints in each frame, with each sample marked with its gesture class. The dataset was categorized into four main groups, which are 18 affix classes, 26 alphabet classes, 35 number classes, and 29 word classes, with sentence data manually segmented into individual word units.

To ensure robust evaluation and prevent overfitting, 5-fold stratified cross-validation was employed instead of a fixed train-validation-test split. This approach partitions the dataset into five equally sized folds while maintaining balanced class distribution within each fold. Each fold is used as the test set exactly once, while the remaining folds are combined to form the training set in each iteration [25]. Each fold was trained for 100 epochs, and final performance metrics were averaged across all folds.

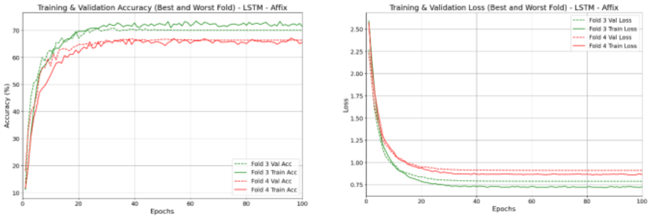
B. Model Performance

The Affix category showed notable variation in performance across folds. Based on 5-fold cross-validation, the best fold (Fold 3) achieved a final training accuracy of 71.42% and a validation accuracy of 70.00%, with the validation loss stabilizing at approximately 0.79. Training loss decreased rapidly during the first 20 epochs and plateaued thereafter, indicating effective learning and convergence. Validation accuracy consistently improved, reaching a stable peak of 70%, reflecting good generalization to the validation set.

In contrast, the worst fold (Fold 4) reached a final training accuracy of 65.52% and a validation accuracy of 66.39%, with the validation loss plateauing at approximately 0.91. Although the early learning phase showed rapid loss reduction and accuracy increase, performance stagnated beyond 66% validation accuracy, suggesting potential overfitting or difficulty in adapting to the validation set.

The loss curves demonstrated rapid decline during the initial epochs, followed by stabilization. Fold 3 maintained consistently lower loss values on both training and validation sets compared to Fold 4, aligning with its superior performance. The accuracy curves revealed a steady increase in validation accuracy for Fold 3, while Fold 4 plateaued and showed greater fluctuation, possibly due to class imbalance or challenging samples.

Overall, convergence was reached around epochs 30 to 40, after which performance stabilized. The best fold (Fold 3) outperformed the worst fold by approximately 4% in validation accuracy. Furthermore, the relatively small gap between training and validation curves in Fold 3 indicated good generalization, whereas the persistent rise in training accuracy with stagnant validation accuracy in Fold 4 pointed to overfitting. Across all folds, the mean validation accuracy was approximately 68.17% with a standard deviation of 1.47%, reflecting stable performance across splits.



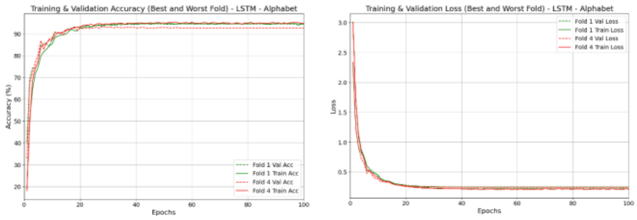
(a) (b)

Figure 6. Training and validation curves for the Affix category, showing (a) accuracy and (b) loss across epochs.

The Alphabet category demonstrated strong and stable performance across folds. In the best-performing fold (Fold 1), the final training accuracy reached 94.48% and the validation accuracy stabilized at 94.49%, with the validation loss plateauing at approximately 0.25. The learning curves showed a rapid reduction in loss and a sharp increase in accuracy during the initial 20 epochs, followed by stable performance across the remaining epochs. Notably, validation accuracy improved consistently throughout training, ultimately maintaining over 94%, reflecting robust generalization.

The worst-performing fold (Fold 4) achieved a final training accuracy of 94.75% and a validation accuracy of 92.74%, with the validation loss stabilizing at around 0.24. Similar to Fold 1, early learning progressed rapidly, but the validation accuracy plateaued earlier and at a slightly lower level, suggesting the presence of more challenging samples or minor data imbalance in the validation split. The loss curves showed parallel trends between training and validation, with Fold 4 maintaining slightly higher validation loss compared to Fold 1.

Both folds exhibited early convergence, with stabilization occurring after approximately 20 to 30 epochs. The gap between training and validation curves remained minimal, indicating good generalization without signs of severe overfitting. Fold 1 consistently outperformed Fold 4 by approximately 1.75% in validation accuracy, highlighting the model's capacity to achieve high performance even in less favorable splits. Overall, the Alphabet category demonstrated high and stable performance, with mean validation accuracy across folds of approximately 93.94% and a standard deviation of 0.63%, indicating robustness across data splits.



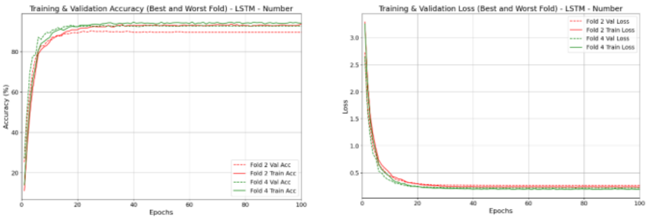
(a) (b)

Figure 7. Training and validation curves for the Alphabet category, showing (a) accuracy and (b) loss across epochs.

The Number category demonstrated consistent and robust performance across folds. In the best-performing fold (Fold 4), the final training accuracy reached 94.14%, with the validation accuracy stabilizing at 92.78% and the validation loss converging to approximately 0.22. The learning curves revealed rapid reductions in both training and validation loss during the first 20 to 30 epochs, followed by a stable plateau. Validation accuracy improved steadily throughout the early epochs and maintained a high level, indicating strong generalization.

In the worst-performing fold (Fold 2), the final training accuracy reached 93.61%, while the validation accuracy plateaued at 89.68%, with the validation loss stabilizing at approximately 0.26. Although Fold 2 followed a similar early learning pattern with rapid gains, the validation accuracy consistently lagged behind Fold 4 by approximately 3%, suggesting minor challenges in the validation split or the presence of harder-to-learn samples.

The accuracy and loss curves for both folds showed early convergence, often in the first 30 epochs. There existed little difference in the training and validation curves, especially in Fold 4, indicating minimal overfitting with good generalization. Fold 4 outperformed Fold 2 overall, indicating that the model could maintain high and consistent performance under good conditions. In all folds, the mean validation accuracy at 91.48% with a standard deviation of 1.18% indicated the model's stability and robustness over data splits.



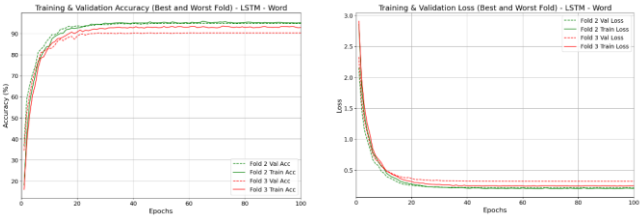
(a) (b)

Figure 8. Training and validation curves for the Number category, showing (a) accuracy and (b) loss across epochs.

The Word category displayed a clear learning pattern with strong performance across folds. The best-performing fold (Fold 2) achieved a final training accuracy of 95.26% and a validation accuracy of 94.83%, with the validation loss stabilizing at approximately 0.22. The learning curves showed a quick drop in loss for the initial 20 to 30 epochs, with a plateau after that, while validation accuracy rose steadily prior to stabilizing above 94%, reflecting very good generalization.

In the poorest performing fold (Fold 3), the training accuracy at the end reached 92.93%, while validation accuracy leveled out at 90.43%, with the validation loss converging at approximately 0.32. While Fold 3 exhibited similar early learning dynamics, the validation accuracy plateaued at a lower level and showed slightly higher fluctuation, possibly due to more challenging samples or imbalance within the validation set.

Both folds showed early convergence, typically within the first 30 epochs, followed by stable performance. The gap between training and validation curves remained minimal, particularly in Fold 2, indicating robust generalization and limited overfitting. Fold 2 outperformed Fold 3 by approximately 4.4% in validation accuracy, highlighting the influence of data splits on performance outcomes. Across all folds, the mean validation accuracy was approximately 92.41% with a standard deviation of 1.65%, reflecting the model's stable behavior across different data partitions.



(a) (b)

Figure 9. Training and validation curves for the Word category, showing (a) accuracy and (b) loss across epochs.

Figure 10 provides a visual summary of the mean test accuracy and its standard deviation across the four SIBI gesture categories. As shown in the figure, the Alphabet category achieved the highest mean test accuracy of 93.94% with minimal variability (±0.63%), followed closely by the Number and Word categories, which achieved mean accuracies of 91.48% (±1.18%) and 92.41% (±1.65%), respectively. In contrast, the Affix category exhibited a notably lower mean test accuracy of 68.17%, with a standard deviation of 1.47%, highlighting a clear performance gap relative to the other categories.

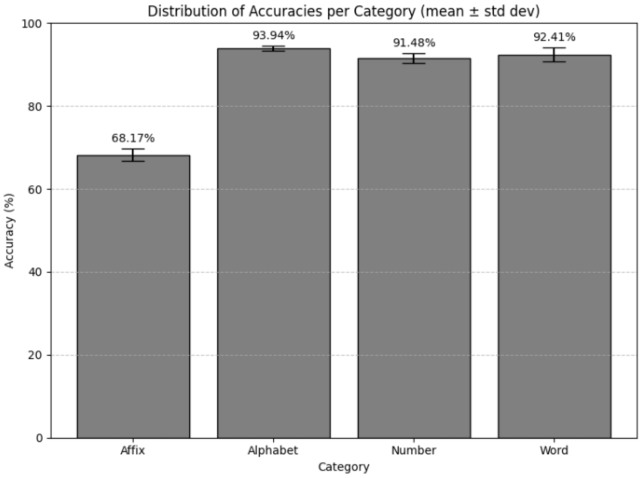


Figure 10. Mean test accuracy and standard deviation across gesture categories (Affix, Alphabet, Number, Word).

The overall performance of the system was evaluated using test accuracy, precision, recall, and F1-score across four categories, namely Affix, Alphabet, Number, and Word, based on five-fold cross-validation. In the Affix category, the model achieved a mean test accuracy of 68.17% with a standard deviation of 1.47%, indicating moderate and somewhat variable performance. The precision, recall, and F1-score averaged 0.6720, 0.6817, and 0.6653, respectively, with low standard deviations (below 0.02), suggesting consistent behavior across folds despite the moderate accuracy level.

Performance in the Alphabet category was particularly robust and consistent. The mean test accuracy was 93.94% with a standard deviation of just 0.63%. Precision, recall, and F1-score were all around 0.94, which indicates excellent consistency and robustness over folds. This performance indicates that the system very effectively recognized alphabet gestures and generalized well over data splits.

Number input category also presented comparable performance, with mean test accuracy at 91.48% with standard deviation 1.18%. Precision, recall, and F1-score averaged 0.9158, 0.9133, and 0.9132, respectively, with tight variation over folds. These performances indicate consistent performance in recognizing numbers with balanced precision and recall leading to high F1-score.

The Word category performed with mean test accuracy at 92.41% with standard deviation of 1.65%. Precision, recall, and F1-score averaged 0.9271, 0.9241, and 0.9239, respectively. Although performance stayed high, there was slightly greater variability compared to Alphabet and Number category, but reasons are unclear, which may be due to higher diversity and complexity of word-level gestures.

However, the overall performance of the system in the Alphabet, Number, and Word category was excellent with high accuracy, precision, recall, and F1-score, and minimal variation over folds. The Affix category presented worse performance, which indicates that closer study of data characteristics, model architecture, or feature space may enhance performance in this class. The balanced precision and recall values over all the classes indicate that the model does not favor or miss any class strongly.

TABLE III
FINAL TEST ACCURACY, PRECISION, RECALL, AND F1-SCORE OF LSTM MODEL ACROSS SIBI GESTURE CATEGORIES

Category	Fold	Test Accuracy	Precision	Recall	F1-score
Affix	1	69.58%	0.6901	0.6958	0.6859
	2	68.19%	0.6655	0.6819	0.6558
	3	70.00%	0.6924	0.7000	0.6906
	4	66.39%	0.6528	0.6639	0.6429
	5	66.67%	0.6592	0.6667	0.6513
	Mean	68.17%	0.6720	0.6817	0.6653
	Std	1,47%	0.0163	0.0147	0.0192
Alphabet	1	94.49%	0.9458	0.9449	0.9442
	2	93.91%	0.9402	0.9390	0.9380
	3	94.39%	0.9440	0.9437	0.9436
	4	92.74%	0.9269	0.9272	0.9267
	5	94.19%	0.9411	0.9417	0.9408
	Mean	93.94%	0.9396	0.9393	0.9387
	Std	0.63%	0.0067	0.0064	0.0064
Number	1	90.64%	0.9069	0.9045	0.9047
	2	89.68%	0.8985	0.8951	0.8947
	3	91.67%	0.9182	0.9154	0.9154
	4	92.78%	0.9286	0.9270	0.9264
	5	92.63%	0.9267	0.9248	0.9250
	Mean	91.48%	0.9158	0.9133	0.9132
	Std	1.18%	0.0116	0.0121	0.0121
Word	1	93.19%	0.9339	0.9319	0.9315
	2	94.83%	0.9496	0.9483	0.9485
	3	90.43%	0.9089	0.9043	0.9041
	4	92.93%	0.9323	0.9293	0.9292
	5	90.68%	0.9105	0.9069	0.9060
	Mean	92.41%	0.9271	0.9241	0.9239
	Std	1.65%	0.0154	0.0165	0.0168

C. Confusion Matrix Analysis

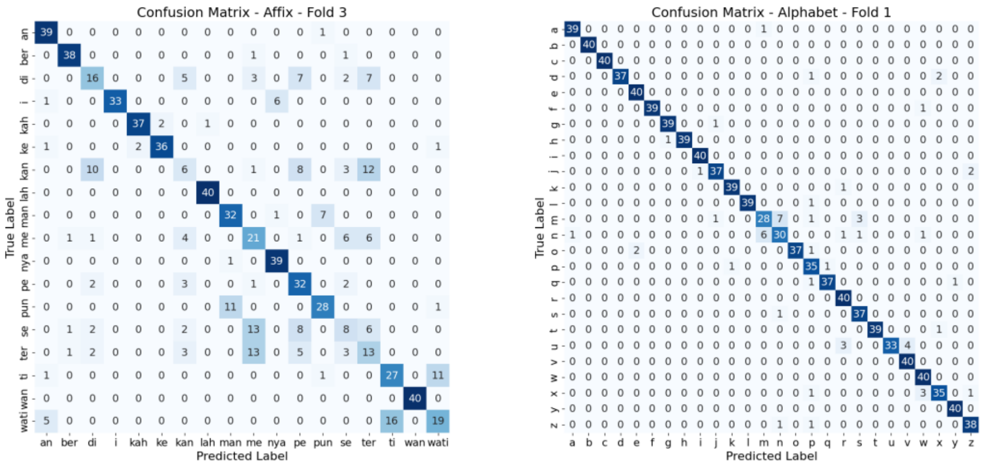
The Affix category performed significantly worse in terms of classification accuracy when compared to the Alphabet, Number, and Word categories, and an examination of the gestures explains why. Numerous affix gestures, for instances like “se” and “me” or “ter” and “me,” vary only with subtle hand shape differences, analogous to thumb location or minimal bending of the fingers, but with similar hand orientation and motion. For an example, both “se” and “me” are done with two hands, but with the thumb in the fist for “me” and the thumb outside the fist for “se.” Similarly, "ter" and "me" vary mostly with the thumb under the index finger holding "ter," with comparable directions of movement. There are other pairs, such as "ti" and "wati" or "pun" and "man," that are very similar in form, with differences occurring only in slight rotations of the hand or in the extent of extended fingers. These minimal visual distinctions make the Affix category problematic not only for learning models, but even for human performances, with resultant misclassifications becoming an expected result.

In contrast, the Alphabet category achieved very high accuracy, with almost all predictions falling perfectly along the diagonal of the confusion matrix. Only minor confusions were observed, such as "m" being predicted as "n". This strong performance is likely because alphabet gestures are characterized by distinct, isolated handshapes with minimal movement, which makes them easier for the model to distinguish.

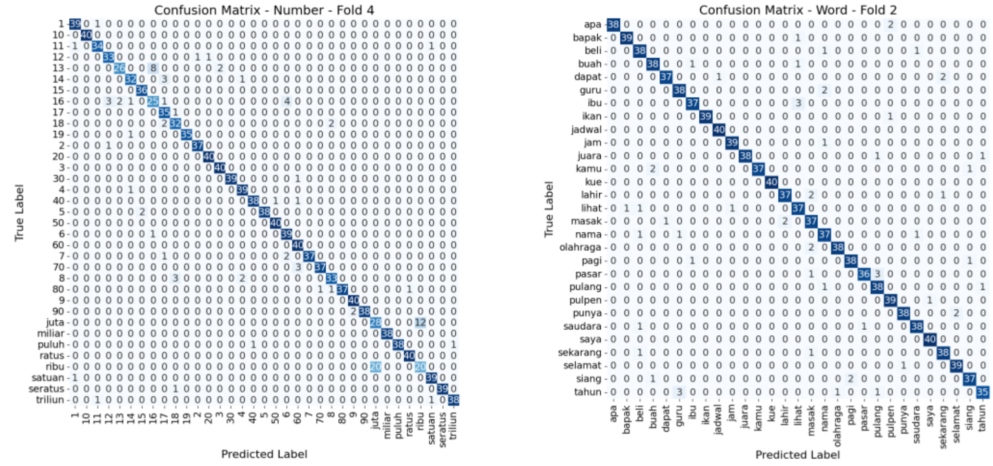
For the Number category, the model also performed well, though some confusion emerged within specific groups. Notably, numbers in the teens, such as "13" and "16" were occasionally misclassified. Similarly, larger magnitude numbers such as "ribu" (thousand) and "juta" (million), showed some overlap. This confusion can be attributed to the fact that numeric gestures often share common patterns or handshapes, particularly in sequential groups, which can make fine-grained distinctions challenging.

The Word category showed strong results with only minor errors. Words like "pasar" (market) and "tahun" (year) were occasionally confused, as were "pulang" (return) and "guru" (teacher). These misclassifications are likely caused by variation in gesture execution or brief occlusion in the video frames. In total, though, word gestures are improved with greater duration, richer context, and more salient motion patterns, which support high accuracy by the model.

Combined, these findings emphasize that the Affix category proves to be the model's greatest challenge due to the category's subtle and visually confusing gestures and high variability among signers, while the Alphabet, Number, and Word categories are classified more strongly. These performance disparities characterize each gesture type's unique visual and temporal features and provide important insights into model strengths and weaknesses. In particular, static gestures like the Alphabet category always have higher accuracy because they present clear, unchanging handshapes, whereas dynamic gestures like the Word and Affix categories add challenges due to movement, temporal variation, and variation among signers. This disparity is especially marked in the Affix category, where gestures often have very similar handshapes and motions, with just subtle variability in the positions of the fingers or slight rotations, making them particularly challenging to separate even with spatial-temporal modeling.



(a) (b)



(c) (d)

Figure 11. Confusion Matrix Visualizations For the Best Fold in (a) Affix, (b) Alphabet, (c) Number, and (d) Word categories

IV. Conclusion

This paper proposes an LSTM-driven hand gesture recognition for the Indonesian Sign Language System (SIBI) for four key gesture types: affix, alphabet, number, and word. The results indicate robust recognition performance with over 91% accuracy on the alphabet, number, and word types, with precision, recall, and F1-scores being high. The affix class performance was the lowest at around 68% due to the increased complexity and visual similarity of affix gestures. Overall, the paper showcases the capability of LSTM models to learn and generalize from sequence keypoint patterns for sign language recognition and presents valuable insights across the four SIBI types.

Some limitations exist for this study, especially the poorer performance when identifying affix signs. This may be a result of limited dataset diversity or difficulties with the design of the model. The system also only deals with manual hand keypoints and lacks the inclusion of non-manual features such as body posture or facial expressions, which restricts its performance when translating full sentences. Future studies need to overcome these limitations by gathering diverse datasets, considering more sophisticated models like attention-based or multimodal models, and evaluating the system for actual time performance to improve the daily communication needs for the Deaf community of Indonesians. Future studies also need validation experiments using new signers to evaluate how well the model generalizes across different signers with different signing styles.

Acknowledgement

The authors would like to sincerely thank the three Deaf participant volunteers for their time and willingness to share their perspectives on the application of sign language on a daily basis. We would also like to sincerely thank the AUDISI Foundation and SLB B Pangudi Luhur for their support and permission for field data collection and observations. We would also like to sincerely thank I Dewa Made Bayu Atmaja for access to the SIBI dataset, which formed an integral part of this research study. Finally, we would like to sincerely thank our families for their constant emotional support and technical feedback throughout this work, and a dedicated colleague who provided much appreciated assistance in the early stages of this research and provided encouragement throughout the research provision.

Bibliography

[1] Badan Pusat Statistik, “Jumlah Penduduk Berumur 5 Tahun ke Atas menurut Kelompok Umur, Daerah Perkotaan/Perdesaan, Jenis Kelamin, dan Tingkat Kesulitan Mendengar, di INDONESIA - Dataset - Long Form Sensus Penduduk 2022,” Badan Pusat Statistik. Accessed: May 02, 2025. [Online]. Available: https://sensus.bps.go.id/topik/tabular/sp2022/145/0/0

[2] N. Napsiah and Y. T. Wijayanti, “Indonesian Society is Not Disabled Friendly?,” *Jurnal Ilmu Sosial*, vol. 22, no. 1, pp. 147–164, Jun. 2023, doi: 10.14710/JIS.22.1.2023.147-164.

[3] M. Nur Iman, “Sign Language and Culture: Understanding Communication in the Deaf Community,” in *Proceeding of the International Conference on Social Sciences and Humanities Innovation*, Asosiasi Peneliti Dan Pengajar Ilmu Sosial Indonesia, 2024, pp. 156–166. [Online]. Available: https://prosiding.appisi.or.id/index.php/ICSSHI

[4] R. S. Fauzi, B. Irmawati, and N. Agitha, “KADARING SIBI (Indonesian Sign System Online Dictionary): Web-based Indonesian Sign System Learning App,” in *Proceedings of the First Mandalika International Multi-Conference on Science and Engineering 2022, MIMSE 2022 (Informatics and Computer Science)*, Atlantis Press International BV, Dec. 2022, pp. 427–436. doi: 10.2991/978-94-6463-084-8_35.

[5] Y. Arief, “Personal Interview with AUDISI Foundation,” Oct. 12, 2024, *Jakarta*.

[6] I. Damayanti and S. H. Purnamasari, “Relationship between Communication Barriers and Stress in Parents with Deaf Children in Elementary Level Special Needs School in Pekanbaru,” *Indonesian Journal of Disability Studies*, vol. 6, no. 1, pp. 14–20, May 2019, doi: 10.21776/UB.IJDS.2019.006.01.2.

[7] S. N. Budiman, S. Lestanti, H. Yuana, and B. N. Awwalin, “SIBI (Sistem Bahasa Isyarat Indonesia) berbasis Machine Learning dan Computer Vision untuk Membantu Komunikasi Tuna Rungu dan Tuna Wicara,” *Jurnal Teknologi dan Manajemen Informatika*, vol.

9, no. 2, pp. 119–128, 2023, Accessed: May 02, 2025. [Online]. Available: http://jurnal.unmer.ac.id/index.php/jtmi

[8] E. Rakun, A. M. Arymurthy, L. Y. Stefanus, A. F. Wicaksono, and I. W. W. Wisesa, "Recognition of Sign Language System for Indonesian Language Using Long Short-Term Memory Neural Networks," *Adv Sci Lett*, vol. 24, no. 2, pp. 999–1004, Mar. 2018, doi: 10.1166/asl.2018.10675.

[9] S. Hidayat, Y. V. Via, and E. P. Mandyartha, "Penerapan Model Hybrid Convolutional Neural Network dan Long Short-Term Memory untuk Pengenalan Real-Time Sistem Isyarat Bahasa Indonesia (SIBI)," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 8, no. 3, p. 1586, Jul. 2024, doi: 10.30865/mib.v8i3.7837.

[10] F. X. L. Riberu, "Sistem Deteksi Simbol pada SIBI (SISTEM ISYARAT BAHASA INDONESIA) Secara Real-Time Menggunakan Mediapipe dan LSTM," Universitas Dinamika, 2023.

[11] I. D. M. B. A. Darmawan *et al.*, "Advancing Total Communication in SIBI: A Proposed Conceptual Framework for Sign Language Translation," in *Proceedings - International Conference on Smart-Green Technology in Electrical and Information Systems, ICSGTEIS*, Institute of Electrical and Electronics Engineers Inc., Nov. 2023, pp. 23–28. doi: 10.1109/ICSGTEIS60500.2023.10424020.

[12] I. D. M. B. A. Darmawan, Linawati, G. Sukadarmika, N. M. A. E. D. Wirastuti, and R. Pulungan, "Temporal Action Segmentation in Sign Language System for Bahasa Indonesia (SIBI) Videos Using Optical Flow-Based Approach," *Jurnal Ilmu Komputer dan Informasi*, vol. 17, no. 2, pp. 195–202, Jun. 2024, doi: 10.21609/jiki.v17i2.1284.

[13] Lembaga Penelitian dan Pengembangan Sistem Isyarat Bahasa Indonesia, "Kamus SIBI." Accessed: May 08, 2025. [Online]. Available: https://pmpk.kemdikbud.go.id/sibi/kosakata/imbuhan

[14] I. D. M. B. A. Darmawan, L. Linawati, G. Sukadarmika, N. M. A. E. D. Wirastuti, and R. Pulungan, "Indonesian Sign Language System (SIBI) Dataset," *Mendeley Data*, vol. 3, Aug. 2024, doi: 10.17632/44PBRBSNKH.3.

[15] Y. Meng, H. Jiang, N. Duan, and H. Wen, "Real-Time Hand Gesture Monitoring Model Based on MediaPipe's Registerable System," *Sensors 2024, Vol. 24, Page 6262*, vol. 24, no. 19, p. 6262, Sep. 2024, doi: 10.3390/S24196262.

[16] I. Galanakis, R. F. Soldatos, N. Karanikolas, A. Voulodimos, I. Voyiatzis, and M. Samarakou, "A MediaPipe Holistic Behavior Classification Model as a Potential Model for Predicting Aggressive Behavior in Individuals with Dementia," *Applied Sciences (Switzerland)*, vol. 14, no. 22, p. 10266, Nov. 2024, doi: 10.3390/APP142210266/S1.

[17] M. S. Jayaprada *et al.*, "Real-Time Hand Gestures Recognition System," *International Journal of Innovative Research in Technology*, vol. 11, no. 11, pp. 4948–4951, 2025, doi: 10.33168/JSMS.2022.0225.

[18] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Networks*, vol. 18, no. 5–6, pp. 602–610, Jul. 2005, doi: 10.1016/J.NEUNET.2005.06.042.

[19] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, Jan. 2014, doi: 10.5555/2627435.2670313.

[20] H. il Lim, "A Study on Dropout Techniques to Reduce Overfitting in Deep Neural Networks," in *Lecture Notes in Electrical Engineering*, Springer, Singapore, Dec. 2020, pp. 133–139. doi: 10.1007/978-981-15-9309-3_20.

[21] PyTorch, "CrossEntropyLoss — PyTorch 2.7 documentation." Accessed: May 05, 2025. [Online]. Available: https://pytorch.org/docs/stable/generated/torch.nn.CrossEntropyLoss.html

[22] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," *Sci Rep*, vol. 14, no. 1, pp. 1–14, Dec. 2024, doi: 10.1038/S41598-024- 66611-y.

[23] C. Miller, T. Portlock, D. M. Nyaga, and J. M. O'Sullivan, "A review of model evaluation metrics for machine learning in genetics and genomics," *Frontiers in Bioinformatics*, vol. 4, p. 1457619, Sep. 2024, doi: 10.3389/FBINF.2024.1457619/XML/NLM.

[24] S. Sathyanarayanan and B. R. Tantri, "Confusion Matrix-Based Performance Evaluation Metrics," *African Journal of Biomedical Research*, vol. 27, no. 4S, pp. 4023–4031, Nov. 2024, doi: 10.53555/AJBR.V27I4S.4345.

[25] I. K. Nti, O. Nyarko-Boateng, and J. Aning, "Performance of Machine Learning Algorithms with Different K Values in K-fold Cross-Validation," *International Journal of Information Technology and Computer Science*, vol. 13, no. 6, pp. 61–71, Dec. 2021, doi: 10.5815/IJITCS.2021.06.05.