Journal of Data Science, Informatics and Computer Engineering (JDSICE),
Vol. 1, Issue 1
September 2026
ISSN: XXXX-YYYY

A Mobile-Based BISINDO Recognition and Communication Support System Using MediaPipe, MLP, and GRU

Adrian Yoga Chrisarianto[1]*, Entin Martiana Kusumaningtyas[2], Aliridho Barakbah[3], Ira Prasetyaningrum[4]
[1]*Associate Degree Program in Computer Science, Electronic Engineering Polytechnic Institute of Surabaya, Surabaya, Indonesia*
**Corresponding author: Adrian Yoga Chrisarianto; email: adrianyogachrisarianto@email.com*

Manuscript received August 14, 2026, revised September 11, 2026, accepted September 22, 2026
doi: zz.yyyyy/xxxxx.pp134-1399

Abstract – *Communication between Indonesian Sign Language users and the general public remains limited because public familiarity with the language is low, while many existing recognition depend on dedicated sensors, run on desktop or web platforms, or recognize only static alphabets. This study develops a landmark-based mobile BISINDO recognition system that supports introductory communication through alphabet recognition, dynamic word recognition, and text-to-speech output. The system provides a mobile-oriented translation workflow using a Streamlit-based interface, camera streaming, prediction stabilization, and text-to-speech output through an external mobile integration bridge. The alphabet dataset contains 13,152 landmark samples, and the word dataset contains 3,600 sequences. In live-camera testing, the alphabet model correctly classified 233 of 260 trials (89.62%), whereas the word model correctly classified 282 of 360 trials (78.33%). Across three users, alphabet accuracy ranged from 84.62% to 100%, and word accuracy ranged from 77.78% to 83.33%. Performance was strongest at 50–80 cm and declined at 110 cm. Sentence construction capability was evaluated by combining recognized alphabet and word outputs into introductory communication sequences. The findings demonstrate the feasibility of an integrated BISINDO recognition workflow for introductory communication scenarios, while future improvements should focus on vocabulary expansion, broader signer-independent evaluation, and mobile deployment optimization.*

CC BY SA

Keywords: *Dynamic gesture recognition, Indonesian Sign Language, landmark-based hand recognition, mobile assistive technology, multilayer perceptron*

I. Introduction

Communication barriers remain a practical problem for Deaf and speech-impaired people when communicating with members of the public who do not understand sign language. The World Health Organization estimates that approximately 1.5 billion people worldwide live with some degree of hearing loss [1]. A survey cited in the accompanying thesis also reported limited public familiarity with Indonesian Sign Language, with only 15% of 100 respondents indicating strong familiarity [2]. Indonesian Sign Language, commonly known as BISINDO, is a visual language used by the Indonesian Deaf community through hand gestures, facial expressions, and body movement. Word signs may vary across regions, while alphabetic finger spelling is comparatively consistent [3]. Mobile devices provide a practical channel for assistive communication. Indonesian telecommunications statistics reported that 68.65% of the population owned a mobile phone in 2024 [4]. A camera-based recognition that does not require specialized gloves or external sensors can therefore reduce adoption barriers, provided that the recognition model is sufficiently accurate, computationally manageable, and robust to signer and environmental variation. Previous studies address only parts of this problem. SETARA used a camera-based Android approach for BISINDO words but emphasized qualitative effectiveness and vocabulary expansion [5]. A flex-sensor glove using k-nearest neighbors achieved 65.38% accuracy but depended on dedicated hardware [6]. U-SIBI provided a web learning environment for SIBI alphabets and vocabulary [7]. A random-forest BISINDO alphabet recognition reported 91.28% accuracy but focused on static alphabet recognition and desktop use [8].

A convolutional neural network web translator recognized SIBI and BISINDO alphabets but did not integrate dynamic BISINDO words into a mobile communication workflow [9].

TABLE I
SYSTEMATIC COMPARISON WITH CLOSELY RELATED SYSTEMS

Study	Scope	Platform and method	Main Difference/ Limitation or distinction
SETAR A [5]	BISIND O words	Android; camera; convolutional network and MediaPipe	Focuses on BISINDO word recognition, but vocabulary coverage and quantitative evaluation remain limited.
Flex glove [6]	SIBI letters and numbers	Sensor glove; k-nearest neighbors	Achieves 65.38% accuracy, but requires additional hardware for gesture acquisition..
U-SIBI [7]	SIBI alphabet and vocabul ary	Web; convolutional network and Flask	Designed mainly as a learning platform and does not provide an integrated BISINDO mobile communication workflow.
Rando m forest [8]	BISIND O alphabet	Desktop; random forest	Provides alphabet classification performance, but only handles static gestures without dynamic words.
Web CNN [9]	SIBI and BISIND O alphabet s	Web; VGG16 convolutional network	Focuses on image-based alphabet recognition without integrated mobile word recognition.
This study	26 BISIND O letters and 18 phrases	MediaPipe, multilayer perceptron, gated recurrent unit, Flutter	Integrates static and dynamic recognition, mobile camera interaction, sentence construction, text-to-speech output, although vocabulary size and signer generalization remain limited.

Study	Scope	Platform and method	Main Difference/ Limitation or distinction
This study	26 BISIND O letters and 18 phrases	MediaPipe, multilayer perceptron, gated recurrent unit, Flutter	Integrates static and dynamic recognition, mobile camera interaction, sentence construction, text-to-speech output, although vocabulary size and signer generalization remain limited.

Recent work has introduced word-level BISINDO datasets and temporal baselines [10], [11], demonstrating that dynamic recognition is an active research direction. However, the supplied literature does not show a single system that combines static finger spelling, dynamic introductory phrases, live mobile access, stabilized output, local history, and speech output. The gap addressed here is therefore primarily a system-integration and practical-evaluation gap rather than a claim of a new learning algorithm.

Unlike previous studies that mainly focus on alphabet recognition or isolated word recognition, this study contributes an integrated BISINDO communication support workflow combining static alphabet recognition, dynamic word recognition, mobile camera interaction, sentence construction, and text-to-speech output. The novelty of this work does not lie in proposing a new recognition algorithm, but in integrating established techniques into a practical BISINDO-oriented assistive communication prototype.

The contributions are threefold. First, the study integrates landmark-based static and temporal recognition in one recognition and communication support system: a multilayer perceptron classifies 26 alphabet gestures, while a gated recurrent unit processes 18 dynamic words and phrases. Second, it evaluates the system through internal validation, live-camera trials, three-user tests, camera-distance tests, and predefined sentence construction. Third, it implements a one-way mobile communication workflow using Flutter, WebView, local SQLite history, and text-to-speech output. Section II describes the proposed system and experiment, Section III presents the results and scientific discussion, and Section IV concludes the paper.

II. PROPOSED SYSTEM AND EXPERIMENTAL METHOD

A. Design and Method Selection

The system avoids direct classification of full image pixels. MediaPipe Hands extracts 21 three-dimensional landmarks from each detected hand [12]. Each detected hand is represented as a set of 21 three-dimensional landmarks, where each landmark consists of normalized coordinates (x_i, y_i, z_i):

$$L = \{ (x_i, y_i, z_i) \mid i = 0, 1, \ldots, 20 \} \quad (1)$$

Therefore, each hand produces 63 coordinate features calculated as:

$$F_\{landmark\} = 21 \times 3 = 63 \quad (2)$$

Landmark-based representations reduce dependence on background pixels and allow spatial relationships to be represented through normalized coordinates, inter-point distances, and orientation features [13]. For alphabet recognition, the landmark representation is combined with geometric descriptors to construct the input vector for the multilayer perceptron. The final feature representation is formulated as:

$$F_\{MLP\} = 63 + (5 + 10 + 5 + 2) = 85 \quad (3)$$

where the additional features consist of fingertip-to-wrist distances, fingertip pair distances, joint angles, and handedness indicators. The multilayer perceptron is selected for static alphabet gestures because its input is a fixed numerical feature vector; previous work has demonstrated the suitability of MediaPipe features with multilayer perceptrons for hand-gesture classification [14].

$$X_GRU \in R\text{^}(30 \times 127) \quad (4)$$

where each sequence consists of 30 frames and each frame contains 127 features, calculated from one hand-count feature and two hand landmark representations:

$$127 = 1 + 63 + 63 \quad (5)$$

The gated recurrent unit is selected for dynamic words because it can model temporal dependencies across landmark sequences while using fewer gates than a long short-term memory network [15]. A majority voting mechanism using a six-frame sliding buffer with a three-vote threshold is applied to reduce frame-level fluctuations [16].

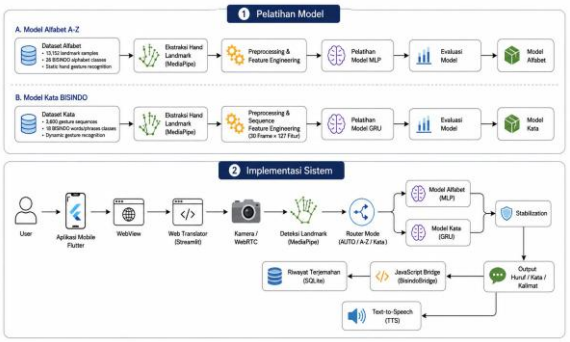

Fig. 1. Overall architecture of the proposed BISINDO recognition and communication support system using MediaPipe, MLP, and GRU.

B. Dataset Acquisition and Representation

The alphabet dataset contains 13,152 landmark samples distributed across 26 classes, with 501–542 samples per class. Each sample represents one static BISINDO alphabet gesture and is converted into an 85-dimensional feature vector consisting of normalized landmark coordinates and geometric descriptors. The dynamic dataset contains 3,600 sequences covering 18 introductory words and phrases, with 200 sequences per class. Each dynamic sequence consists of 30 frames, where each frame is represented by 127 temporal landmark features.

The datasets were collected from two voluntary signers for both alphabet and dynamic word recognition. Participants A and B contributed to the dataset collection, and both participants were also included in the real-time evaluation process, while Participant C was an additional unseen user used for generalization testing. Before data collection, participants were informed about the purpose of the study and agreed to participate voluntarily.

The recording process was conducted under several practical environmental conditions, including a laboratory environment with stable lighting, a boarding-room environment with moderate lighting, and a coffee-shop environment with less stable lighting. The camera was positioned at several distances representing near, medium, and relatively far usage conditions. Each recorded sample was manually checked to ensure consistency between the performed gesture and its corresponding BISINDO label before being used for training. The dynamic word dataset was divided using an 80/20 stratified split, resulting in 2,880 training sequences and 720 validation sequences, with approximately 40 validation sequences allocated for each class.

The internal dataset partition was used for model training and validation, while generalization capability was further examined through live-camera evaluation involving an additional unseen participant. However, the current study does not represent a fully signer-independent training-validation protocol because two evaluation participants were also involved in dataset collection. Therefore, the remaining performance gap between validation accuracy and real-world testing may

be affected by signer variation, gesture execution style, and environmental conditions. Future work should adopt a larger subject-independent evaluation protocol with separated training and testing signers.

Although Participant C represents an unseen signer evaluation, the current experiment does not represent a fully signer-independent protocol because Participants A and B contributed to dataset collection and were also included in the evaluation process. Therefore, the obtained results should be interpreted as prototype-level validation under the tested conditions rather than as a general representation of BISINDO recognition performance across diverse users. Future work should involve a larger number of independent signers with separated training and testing participants to evaluate broader generalization capability.

C. Training Configuration and Reproducibility

Both training curves in the thesis extend to 100 epochs. The word pipeline uses a 30-frame by 127-feature input, while the alphabet pipeline uses a dual-hand feature vector. The final training configuration used in this study is summarized in Table II to support reproducibility.

TABLE II
MODEL AND TRAINING CONFIGURATION

Parameter	Alphabet Model	Word Model
Input	85-dimensional engineered feature vector	30×127 landmark sequence
Feature extraction	MediaPipe + geometric features	MediaPipe temporal landmarks
Classes	26 alphabets	18 words/phrases
Model	MLP	GRU
Architecture	MLP with Dense layers (256, 128, 64) and 26-class Softmax output.	GRU(64) → Dense(32) → Dense(18 Softmax)
Activation	ReLU + Softmax	tanh + ReLU + Softmax
Dropout	0.35/0.30/0.20	0.20
Optimizer	Adam	Adam
Learning rate	0.001	0.001
Epoch	100	100
Batch size	32	32
Loss	Categorical CE + LS 0.05	Categorical CE + LS 0.05
Split	80/20 stratified	80/20 stratified sequence split
Evaluation	Accuracy, F1, confusion matrix	Accuracy from validation and real-time testing

Camera distance evaluation was conducted at 50 cm, 80 cm, and 110 cm. Ten predefined introductory sentences were tested until the target sentence was successfully produced. The proposed system was implemented as a mobile-based application; however, Detailed analysis of application-level performance, such as execution latency, frame rate stability, and hardware resource consumption, is outside the scope of this study. Therefore, the current evaluation focuses on recognition accuracy and functional camera-based interaction rather than detailed mobile hardware performance analysis.

D. Mobile Integration and Evaluation Protocol

The recognition system is executed through a Streamlit web interface displayed inside a Flutter WebView. Camera frames are acquired through the web recognition, processed by MediaPipe, routed to alphabet, word, or automatic mode, stabilized, and returned as letters, words, or sentences. The system provides a JavaScript communication bridge for transferring recognized text to a mobile host application, where additional storage and speech services can be implemented [17]–[19]. This is a one-way recognition workflow that converts recognized signs into text and speech output; it does not support reverse translation from speech or text into signs.

Internal evaluation of the alphabet model used accuracy, loss, macro F1-score, and confusion matrix analysis, while the word model evaluation was based on validation accuracy obtained during model evaluation and real-time camera testing. Live evaluation was conducted using a smartphone equipped with a Snapdragon 6s Gen 3 processor, 8 GB RAM, Android 16 operating system, and dual cameras (50 MP main camera and 8 MP front camera). Alphabet testing consisted of 10 trials for each of 26 classes (260 trials), while word testing consisted of 20 trials for each of 18 classes (360 trials). User variation was evaluated using three participants, including two participants involved in dataset collection and one additional unseen participant. Camera distance experiments were conducted at 50 cm,
80 cm, and 110 cm. Ten predefined introductory sentences were tested to evaluate sentence construction capability, where each sentence was attempted until the expected output was produced. Although the system was evaluated through real-time camera interaction,

comprehensive device-level benchmarking, including application latency, frame rate, and resource consumption, was not conducted.

III. RESULTS AND DISCUSSION

The testing was conducted using a smartphone equipped with a Snapdragon 6s Gen 3 chipset (8 GB of RAM).

A. Internal Model Behaviour

For both models, the training history reports final training accuracy of approximately 0.86–0.87 and validation accuracy of approximately 0.90–0.91. Training loss decreased from about 1.60 to 0.76, and validation loss decreased from about 1.07 to 0.63. The alphabet confusion matrix is predominantly diagonal, with errors concentrated among visually similar hand configurations. The thesis identifies confusions involving C and L and additional pairs such as U–C, P–K, H–D, B–F, and T–K.

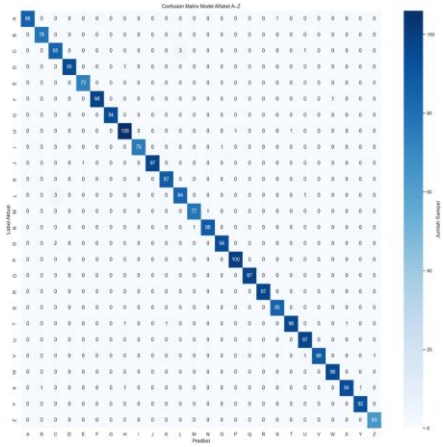

Fig. 2. Confusion matrix from the validation test on the A–Z Alphabet (MLP) model.

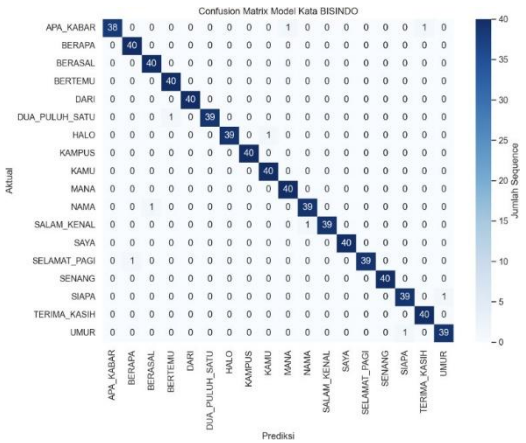

Fig. 3.Validation confusion matrix for the 18-word BISINDO GRU model

The word confusion matrix shows a strong diagonal pattern, indicating that most BISINDO word classes were correctly classified by the GRU model. The validation results demonstrate that the model successfully learned temporal patterns from the gesture sequences under the prepared dataset conditions. However, these internal validation results should be interpreted carefully because they represent performance on the collected dataset and do not fully represent recognition performance for unseen users and uncontrolled environments. The narrative reports approximately 0.90–0.91 validation accuracy, while another passage states that, and the displayed matrix contains counts close to 40 per class rather than the 20 stated in the text. Consequently, the validation results should be interpreted with careful consideration because consistency between reported metrics and exported evaluation artifacts is required to ensure reproducibility. The present discussion therefore treats the live-camera results as the more relevant deployment evidence rather than presenting the internal matrix as proof of signer-independent generalization.

B. Live-Camera Performance

TABLE III
LIVE-CAMERA Results

Test	Alphabet model	Word model
Overall manual test	233/260 (89.62%)	282/360 (78.33%)
User A	100.00%	83.33%
User B	84.62%	83.33%
User C	84.62%	77.78%
50 cm	100.00%	80.00%
80 cm	100.00%	80.00%
110 cm	84.61%	60.00%

The alphabet model achieved 89.62% across 260 live trials, while the dynamic word model achieved 78.33% across 360 trials. The lower word accuracy is consistent with the greater sensitivity of temporal gestures to movement speed, repetition timing, hand trajectory, and landmark interruption. Multi-user results also show that performance changes across signers. These tests are useful practical checks, but three participants are insufficient for a strong subject-independent claim.

Camera distance affected both models. Alphabet accuracy remained 100% at 50 and 80 cm and fell to 84.61% at 110 cm. Word accuracy was 80% at 50 and 80 cm and fell to 60% at 110 cm. As the hand occupies fewer pixels at longer distances, MediaPipe landmark localization becomes less stable. The system should therefore communicate a recommended operating range of 50–80 cm and should not present the current results as

distance-invariant performance.

C. Conversational Function and Practical Contribution

All ten predefined introductory sentences were eventually produced in 17 total attempts, corresponding to 1.7 attempts per sentence. Six sentences succeeded on the first attempt, giving a first-attempt sentence success rate of 60%. The remaining sentences required two to four attempts, particularly when switching between word mode and alphabetic finger spelling. The appropriate interpretation is therefore 100% eventual completion on a closed set of ten target sentences, not 100% open-conversation accuracy.

The practical contribution is the integration of alphabet and dynamic-word recognition with saved history and speech output. Nevertheless, the current 18-word vocabulary corresponds to only 36%–56.25% of the class counts in three external word-level BISINDO resources discussed in the thesis [10], [11]. These percentages describe class-count coverage only and do not establish accuracy superiority. Direct comparison with the 91.28% random-forest alphabet result [8] is also inappropriate because the datasets, users, and evaluation protocols differ. The present system should therefore be positioned as an integrated prototype rather than a state-of-the-art accuracy claim.

D. Explanation of the Validation-to-Deployment Gap

The decrease from approximately 0.90–0.91 internal validation accuracy to 78.33% live word accuracy indicates a domain shift between curated sequences and uncontrolled camera input. The thesis identifies lighting, camera angle, user anatomy, movement speed, gesture consistency, and landmark stability as plausible contributors. A further concern is that the dataset split is not documented as signer-independent. A rigorous revision should use a subject-independent protocol, such as leave-one-signer-out validation or a held-out signer test set, and report macro precision, recall, F1-score, confidence intervals, and per-class live errors. An ablation comparing raw landmarks, engineered features, and prediction stabilization would also clarify which components produce the observed gains.

IV. CONCLUSION

This study presents a landmark-based mobile BISINDO recognition system that combines a multilayer perceptron for 26 static alphabet classes and a gated recurrent unit for 18 dynamic introductory words and phrases. Live-camera evaluation achieved 89.62% accuracy for the alphabet model and 78.33% accuracy for the word model. The system achieved its best operating performance at camera distances of 50–80 cm, and ten predefined introductory sentences were successfully constructed by combining recognized alphabet and word outputs. These findings demonstrate the preliminary feasibility of an integrated mobile-oriented BISINDO recognition workflow for limited introductory communication scenarios. However, the current system should be considered as a communication support prototype rather than a complete BISINDO translation system, since vocabulary coverage and signer-independent evaluation remain limited.

Appendix

A video documenting the results of the sentence construction and real-time detection tests can be accessed via the additional repository link at: https://drive.google.com/drive/folders/19BiXZH5LSocU3FtSEIainsH8W7FdCyp0:

Acknowledgements

The author would like to thank the Department of Information and Computer Engineering at the Surabaya State Electronics Polytechnic (PENS) for providing research facilities. Special thanks are also extended to Ms. Entin Martiana K., Mr. Ali Ridho Barakbah, and Ms. Ira Prasetyaningrum, the academic advisors who provided comprehensive guidance throughout the process, from the design to the evaluation of this system.

References

[1] World Health Organization, World Report on Hearing. Geneva, Switzerland: World Health Organization, 2021.
[2] "Perancangan Media Pembelajaran Interaktif Pengenalan Dasar-Dasar BISINDO untuk Masyarakat Non-Tunarungu," Telkom University Repository, 2025.
[3] Direktorat SMP Kemendikbudristek, "Mengenal Bahasa Isyarat di Indonesia: SIBI dan BISINDO," 2023.
[4] Badan Pusat Statistik, Statistik Telekomunikasi Indonesia 2024. Jakarta, Indonesia: BPS, 2025.
[5] J. T. Sihotang and A. N. A. Adena, "SETARA (Aplikasi Penerjemah Bahasa Isyarat) sebagai Solusi Disekuilibrium Komunikasi dan Pendidikan di Kota Semarang," Journal for Energetic Youngsters, vol. 2, no. 1, pp. 27–45, 2024.
[6] R. A. Zakaria, A. Firmansyah, and Y. A. Prabowo, "Rancang Bangun Flex Sensor Gloves untuk Penerjemah Bahasa Isyarat Menggunakan K-Nearest Neighbors," in Seminar Nasional Sains dan Teknologi Terapan VII, Surabaya, Indonesia, 2019.
[7] N. A. Dzikri and D. A. Rosdiana, "Aplikasi Website Penerjemah Bahasa Isyarat sebagai Media Pembelajaran SLB Negeri Cicendo Kota Bandung," Jurnal Ilmiah Rekayasa dan Manajemen Sistem Informasi, vol. 10, no. 1, pp. 42–50, 2025.
[8] A. Hadi and L. L. Van FC, Rancang Bangun Aplikasi Penerjemah BISINDO Menggunakan Metode Random Forest Classifier, undergraduate thesis, Universitas Lancang Kuning, Pekanbaru, Indonesia, 2023.
[9] J. A. Jong, Aplikasi Penerjemah Bahasa Isyarat Menggunakan Convolutional Neural Network Berbasis Web, internship report, Universitas Katolik De La Salle Manado, Manado, Indonesia, 2024.
[10] G. O. Kindy, G. Leonali, and H. Lucky, "Word-Level BISINDO: A Novel Video Indonesian Sign Language Dataset

and Baseline Methods," Procedia Computer Science, vol. 269, pp. 249–258, 2025, doi: 10.1016/j.procs.2025.08.277.
[11] M. D. Kautsar, A. M. Hariono, and R. Akmal, "Attention vs LSTM: Improving Word-Level BISINDO Recognition," arXiv:2409.01975, 2024.
[12] F. Zhang, V. Bazarevsky, A. Vakunov, A. Tkachenka, G. Sung, C.-L. Chang, and M. Grundmann, "MediaPipe Hands: On-Device Real-Time Hand Tracking," arXiv:2006.10214, 2020.
[13] J. Shin, A. S. M. Miah, M. H. Kabir, M. A. Rahim, and A. A. Shiam, "A Methodological and Structural Review of Hand Gesture Recognition Across Diverse Data Modalities," IEEE Access, vol. 12, pp. 142606–142611, 2024.
[14] S. Barman and S. Majumdar, "Person-Independent Hand Gesture Recognition Using MediaPipe and Multi-Layer Perceptron," in Smart Systems and Wireless Communication, vol. 433. Singapore: Springer, 2025, doi: 10.1007/978-981-96-1348-9_1.
[15] B. Subramanian, B. Olimov, S. M. Naik, S. Kim, K.-H. Park, and J. Kim, "An Integrated MediaPipe-Optimized GRU Model for Indian Sign Language Recognition," Scientific Reports, vol. 12, Art. no. 11964, 2022.
[16] M. F. Wahid, R. Tafreshi, and R. Langari, "A Multi-Window Majority Voting Strategy to Improve Hand Gesture Recognition Accuracies Using Electromyography Signal," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 28, no. 2, pp. 427–436, 2020, doi: 10.1109/TNSRE.2019.2961706.
[17] Flutter, "Flutter Documentation," 2023.
[18] flutter_tts, "Text to Speech Plugin for Flutter," 2023.
[19] SQLite, "SQLite Documentation," 2023.