Scientific Journal of Informatics
Vol. 13, No. 2, May 2026
p-ISSN 2407-7658 http://journal.unnes.ac.id/nju/index.php/sji e-ISSN 2460-0040

Enhancing Generalization of BISINDO Alphabet Recognition Using Anatomy-Aware Realistic 3D Keypoint Augmentation

Muhammad Fa'iz Alfarisi[1*], Denis Eka Cahyani[2], Samsul Setumin[3]
[1]Faculty of Mathematics and Natural Sciences, Universitas Negeri Malang, Indonesia
[2]Faculty of Mathematics and Natural Sciences, Universitas Negeri Malang, Indonesia
[3]Departement of Electrical Engineering, Universiti Teknologi MARA, Malaysia

Abstract.
Purpose: Sign language recognition systems based on 3D hand keypoints frequently experience generalization issues when trained on limited and homogeneous datasets, particularly under single-subject data collection settings. In BISINDO alphabet recognition, this limitation often leads to significant performance degradation when models are applied to unseen users or different acquisition devices. This study aims to improve cross-domain generalization of BISINDO alphabet recognition models by introducing realistic feature-level augmentation applied directly to 3D hand keypoints.
Methods: A realistic 3D keypoint augmentation framework was proposed, consisting of Gaussian Jitter, Anisotropic Scale, Bone Length Scale, and Depth & Tilt Jitter to simulate sensor noise, anatomical variability, and viewpoint changes. Hand keypoints were extracted using MediaPipe Hands and classified using a multilayer perception (MLP). Model performance was evaluated through k-fold cross-validation on a single-subject internal dataset and cross-domain testing on an external dataset involving unseen subjects and different acquisition devices.
Results: The experimental results indicate that the proposed augmentation strategy substantially improves generalization performance without degrading in-domain accuracy. The cross-domain F1-score increased from 67.20% in the baseline model to 89.55% after applying realistic 3D keypoint augmentation, while performance variability across validation folds was also reduced, indicating more stable learning behavior.
Novelty: This work highlights that controlled geometric manipulation at the 3D keypoint level provides an effective and computationally efficient approach to mitigating overfitting in low-resource BISINDO recognition scenarios. By focusing on feature-level augmentation rather than image-based transformations or algorithm replacement, this study offers a practical strategy for enhancing robustness in real-world sign language recognition systems.

Keywords: BISINDO, Data augmentation, 3D keypoints, Generalization, Cross-domain evaluation
Received February 2026 / **Revised** March 2026 / **Accepted** March 2026

INTRODUCTION

Sign language functions as the primary medium of communication within the Deaf community, relying on structured visual gestures to convey linguistic meaning. In Indonesia, Bahasa Isyarat Indonesia (BISINDO) is widely used in daily interactions among Deaf individuals. However, sign language proficiency among the general population remains limited, which continues to create communication barriers in social, educational, and public service contexts [1] In practice, communication between Deaf and hearing communities often depends on human sign language interpreters. While effective, the availability of professional interpreters is limited and frequently associated with high operational costs, making large-scale and continuous deployment difficult [2]. These constraints have motivated increasing interest in computer vision–based automatic sign language recognition systems as a technological alternative to support more inclusive and scalable communication [3].

Recent developments in sign language recognition have been dominated by deep learning–based approaches, particularly pixel-level methods that operate directly on image data. Convolutional Neural Networks (CNNs) and object detection architectures have shown strong performance in controlled experimental settings [4], [5]. Despite their success, pixel-based models are inherently sensitive to visual variability. Changes in illumination, background complexity, camera quality, and skin tone often lead to performance degradation when such models are deployed outside laboratory conditions [6]. This limitation arises because pixel-level approaches tend to exploit appearance-based patterns rather than explicitly

*Corresponding author.
Email addresses: muhammad.faiz.2203126@students.um.ac.id (Alfarisi)
DOI: 10.15294/sji.v12i1.42221

modeling the geometric structure of hand gestures [7]. As a result, high accuracy reported in controlled environments does not always translate into reliable performance under real-world conditions.

To address these limitations, recent studies have increasingly shifted toward keypoint-based representations that model hand gestures as structured geometric configurations [8], [9], [10], [11]. Frameworks such as MediaPipe Hands enable real-time extraction of three-dimensional (3D) hand keypoints with relatively low computational overhead, making them suitable for mobile and edge-device deployment [12]. By representing gestures as coordinate-based vectors, keypoint-based approaches are less dependent on background appearance and lighting conditions, while significantly reducing data dimensionality [13]. In the context of BISINDO recognition, several studies have demonstrated that keypoint-based models provide competitive accuracy and improved robustness compared to pixel-based methods, particularly for real-time and resource-constrained applications [14].

Despite these advantages, keypoint-based sign language recognition models are not immune to generalization problems. A major challenge lies in the limited availability of diverse training data. Collecting large-scale BISINDO datasets that involve multiple subjects with varying hand anatomies and recording conditions is often constrained by logistical complexity and acquisition costs [15]. Consequently, many studies rely on datasets collected from a single subject or a small number of participants. Under such conditions, neural network models are prone to overfitting, learning subject-specific characteristics such as hand size or finger proportions rather than subject-invariant geometric patterns [16]. While this issue may be masked by high internal evaluation scores, it becomes evident during cross-domain testing, where models are evaluated on unseen subjects or different acquisition devices, often resulting in a significant drop in performance [17], [18].

Data augmentation is a widely adopted strategy to mitigate overfitting under limited data conditions by increasing data diversity without additional data collection [19], [20]. In keypoint-based recognition tasks, spatial augmentation has been shown to improve robustness by introducing controlled geometric variations during training. Evidence from skeleton-based human action recognition indicates that techniques such as noise injection and geometric scaling can enhance generalization by encouraging models to learn invariant structural relationships rather than memorizing exact joint coordinates [21], [22], [23]. However, augmentation strategies designed for full-body motion cannot be directly transferred to hand sign language recognition. Hand gestures exhibit higher spatial complexity, where subtle variations in finger configuration may alter semantic meaning. Previous studies have reported that inappropriate augmentation, such as aggressive cropping, can degrade performance by removing critical joint information [23]. This highlights the need for augmentation strategies that preserve anatomical plausibility and semantic consistency, particularly for finger-level representations [24].

Recent work in pose-based and sign language recognition suggests that feature-level geometric transformations, including noise injection, scaling, and rotation, offer a more stable alternative to image-level augmentation when working with keypoint data [25]. Unlike pixel-based manipulation, these transformations operate directly on coordinate representations, allowing precise control over spatial variation without disrupting hand topology. Prior studies also emphasize that such augmentation is most effective when parameter ranges are carefully constrained to reflect realistic anatomical variation, especially in low-resource settings [26], [27]. These findings indicate that well-designed geometric augmentation at the 3D keypoint level has the potential to improve generalization while maintaining the semantic integrity of hand gestures.

Based on these considerations, this study focuses on improving generalization in BISINDO alphabet recognition under limited data conditions through realistic feature-level 3D keypoint augmentation. Four complementary augmentation techniques are applied: Gaussian Jitter to simulate sensor noise and natural hand tremors, Anisotropic Scale to model variations in hand proportions, Bone Length Scale to account for inter-individual differences in finger anatomy, and Depth & Tilt Jitter to represent changes in camera distance and viewpoint. The effectiveness of these augmentations is evaluated using a Multi-Layer Perceptron (MLP) model trained on a single-subject dataset and tested under cross-domain conditions involving unseen subjects and different acquisition devices. The main contribution of this work lies in demonstrating that controlled geometric manipulation at the 3D keypoint level can serve as an effective and computationally efficient strategy for mitigating overfitting and enhancing cross-domain robustness in low-resource BISINDO recognition scenarios.

METHODS

This study adopts an experimental research design to evaluate the effectiveness of realistic 3D keypoint–based data augmentation in improving the generalization capability of BISINDO alphabet recognition models. The overall workflow follows a structured machine learning lifecycle consisting of data preparation, feature extraction and normalization, feature-level data augmentation, model training, and performance evaluation. [28], [29]. To ensure a fair and interpretable comparison, two experimental configurations are constructed: (1) a baseline model trained without augmentation and (2) a proposed model trained with realistic 3D keypoint augmentation. Both configurations employ identical feature representations, model architecture, and training protocols. This design allows the observed performance differences to be attributed specifically to the effect of the augmentation strategy rather than architectural or optimization factors. The overall research workflow is illustrated in Figure 1.

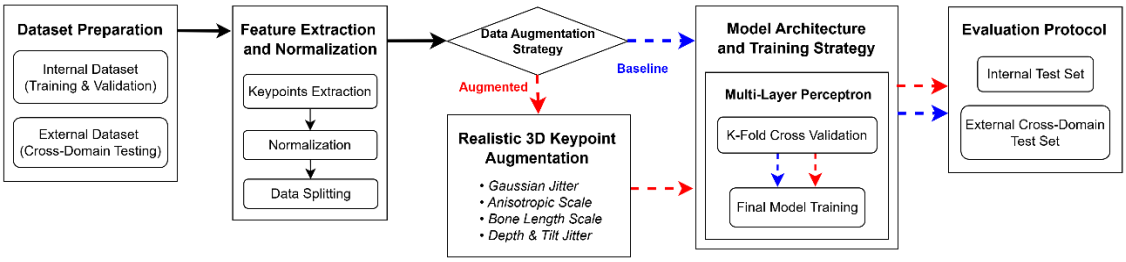

Figure 1. Research flow

Dataset preparation

Two datasets with distinct acquisition characteristics are utilized to assess both in-domain performance and cross-domain generalization. The internal dataset serves as the sole source for model training and validation, while the external dataset is reserved exclusively for cross-domain evaluation. This strict separation ensures that no samples from unseen subjects or devices influence the training process.

Internal dataset

The internal dataset used in this study comes from the Indonesian-Sign-Language-BISINDO-Hand-Sign-Detection-Dataset [30], which is publicly available on GitHub and was accessed in September 2025. In total, the dataset contains 520 images in JPG format, covering 26 BISINDO alphabet classes (A–Z), where each class consists of 20 samples, so the distribution is evenly balanced.

All images were captured using a smartphone camera (Redmi Note 10 Pro) and processed with a simple acquisition setup using Python and OpenCV. During data collection, the same recording setup was consistently applied, including factors such as lighting conditions and the distance between the hand and the camera. Each sample is a standalone image representing a static hand gesture, meaning it is not extracted from a video but directly taken as an individual frame for each alphabet.

Although the dataset is class-balanced, it was collected from a single subject, resulting in limited variability in hand shape, gesture execution, and camera interaction. Such homogeneity is commonly observed in small-scale datasets and may lead to models learning subject-specific characteristics rather than subject-invariant geometric patterns. Consequently, while the dataset is suitable for evaluating in-domain recognition performance, it presents limitations for assessing generalization across different users. Visual examples of the internal dataset are presented in Figure 2.

Figure 2. Internal dataset sample

External dataset

To evaluate the generalization capability of the model, an external dataset was collected independently and used exclusively for testing. The dataset consists of 520 images covering 26 BISINDO alphabet classes (A–Z). In contrast to the internal dataset, this dataset was designed to introduce additional variability, both in terms of subject characteristics and acquisition devices. Data collection was performed using two devices: a smartphone camera (Redmi Note 13 Pro 5G) and a laptop webcam (MSI GF63). The smartphone images were captured in portrait orientation, while the webcam images were recorded in landscape orientation, resulting in differences in framing and sensor characteristics. All images were captured under indoor conditions with natural lighting and relatively simple backgrounds to maintain consistency while still allowing moderate variation in acquisition conditions.

The dataset involves three subjects with different hand characteristics. Subject 1 and Subject 2 have medium-sized hands, while Subject 3 has relatively larger hand proportions, introducing natural anatomical variation across samples. Furthermore, images were captured at varying distances from the camera, leading to differences in hand scale and perspective. Each sample represents a static BISINDO alphabet gesture captured under controlled conditions.

The distribution of samples across subjects and devices was intentionally arranged to introduce variation in acquisition conditions. Subject 1 contributed data from both devices, with five samples per class captured using the smartphone and five using the webcam. Subject 2 contributed samples exclusively from the webcam, while Subject 3 contributed samples exclusively from the smartphone, each providing five samples per class. This setup was primarily intended to increase variation in capture conditions, particularly in terms of device usage and recording perspective. Visual examples of the external dataset are presented in Figure 3.

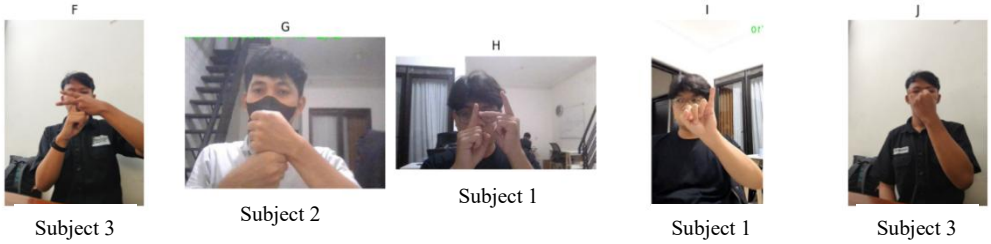

Figure 3. External dataset sample

Feature extraction and normalization

Hand gesture features are extracted using the MediaPipe Hands framework, which estimates three-dimensional (3D) hand keypoints in real time [12]. For each detected hand, MediaPipe Hands provides 21 anatomical keypoints, each represented by spatial coordinates (x,y,z). In practice, BISINDO alphabet gestures may involve either one or two hands, resulting in variability in the number of detected keypoints across samples. To maintain a fixed-length input representation required by the classification model, a zero-padding strategy is applied. When only one hand is detected, the missing hand's keypoints are replaced with zero-valued coordinates. This ensures that every sample is represented by a consistent feature vector without introducing artificial geometric distortion.

After keypoint extraction, the coordinates from both hands are concatenated into a single feature vector. Each hand contributes 21 keypoints with three spatial dimensions, resulting in a total of 126 features per sample ($21\times3\times2$). This compact representation preserves the geometric structure of hand gestures while significantly reducing input dimensionality compared to pixel-based representations

To reduce sensitivity to absolute hand position, scale, and camera distance, geometric normalization is applied prior to model training. First, translation normalization is performed by shifting all keypoints such that the wrist joint serves as the coordinate origin. This step removes positional bias and ensures invariance to hand location within the image frame.

Second, scale normalization is applied to standardize hand size across samples. Following established practice in keypoint-based pose analysis [31], all coordinates are divided by a reference bone length defined as the Euclidean distance between the wrist and the index finger metacarpophalangeal (MCP) joint. This

reference is chosen because it provides a stable anatomical measure that is minimally affected by finger articulation. As a result, the normalized features primarily encode relative geometric relationships rather than absolute spatial measurements.

Realistic 3D keypoint augmentation

All augmentation techniques are applied exclusively to the training subset within each cross-validation fold, while validation and test data remain unchanged. This constraint is imposed to prevent data leakage and to ensure that performance metrics accurately reflect generalization rather than memorization of augmented samples.

Let $P = \{p_i\}_{i=1}^{N}$ denote the set of normalized 3D hand keypoints, where each keypoint is represented as $p_i = (x_i, y_i, z_i)$ and N = 21 for each hand. The proposed augmentation framework introduces four complementary geometric transformations, each designed to simulate a specific source of real-world variability commonly encountered in sign language acquisition.

Figure 4 presents representative examples of geometric variations applied to 3D hand keypoints before and after augmentation. The visualization illustrates how each transformation introduces controlled spatial variability while preserving the overall hand structure and semantic meaning of the gesture.

The applied augmentations are not intended to produce arbitrary distortions, but rather to reflect variations that commonly occur in real-world sign language acquisition, including sensor noise, differences in hand anatomy across users, and changes in camera viewpoint. Each technique modifies the keypoint configuration in a specific manner: Gaussian Jitter introduces small positional perturbations, Anisotropic Scale alters spatial proportions, Bone Length Scale adjusts inter-joint distances, and Depth & Tilt Jitter simulates perspective variation. As shown in Figure 4, these transformations remain anatomically plausible and maintain the relative topology of the hand. The following subsections describe each augmentation technique along with its mathematical formulation and design rationale.

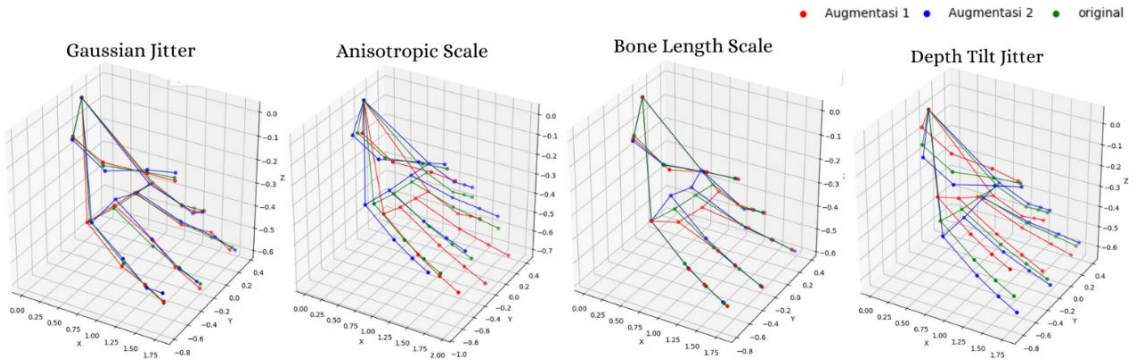

Figure 4. Visualization of realistic 3D keypoint augmentation applied to a BISINDO hand gesture

Gaussian Jitter

Gaussian Jitter is introduced to simulate small positional perturbations arising from sensor noise and natural hand tremors. Each keypoint coordinate is independently perturbed by additive Gaussian noise as shown in Equation (1).

$$p_i' = p_i + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2) \tag{1}$$

where σ controls the noise magnitude. The value of σ is empirically constrained to ensure that the resulting perturbations remain within anatomically plausible bounds. This augmentation encourages the model to focus on stable geometric relationships rather than memorizing exact joint coordinates, a property shown to improve generalization in skeleton-based recognition tasks [23].

Anisotropic Scale

Anisotropic Scale models variability in hand proportions by applying non-uniform scaling along each spatial axis as shown in Equation (2).

$$p_i' = Sp_i, \qquad S = \begin{bmatrix} s_x & 0 & 0 \\ 0 & s_y & 0 \\ 0 & 0 & s_z \end{bmatrix} \tag{2}$$

where s_x, s_y, and s_z are independently sampled scaling factors. Unlike isotropic scaling, which uniformly resizes the hand, anisotropic scaling allows differential deformation along spatial axes. This transformation reflects both natural anatomical variation among individuals and mild perspective distortion caused by camera orientation, while preserving the overall topology of the hand structure.

Bone Length Scale

Bone Length Scale explicitly targets inter-individual differences in finger anatomy by modifying relative bone lengths within the hand kinematic tree. Let p_j and p_k denote a parent–child joint pair, the scaled joint position is computed, as shown in Equation (3).

$$p_k' = p_j + \alpha(p_k - p_j) \tag{3}$$

where α is a scaling factor applied to the bone vector. This operation preserves joint connectivity while introducing realistic variation in finger length. Compared to global scaling, this localized transformation more accurately reflects biological differences across users, making it particularly relevant for subject-invariant hand gesture modeling.

Depth & Tilt Jitter

Depth & Tilt Jitter is introduced to simulate variations in camera distance and viewpoint. Depth variation is modeled by applying translation along the z-axis, as shown in Equation (4).

$$p_i' = p_i + (0,0,\Delta z) \tag{4}$$

where Δz represents a small depth displacement. Additionally, viewpoint changes are modeled using small-angle rotations around the horizontal axis, as shown in Equation (5).

$$p_i' = R_x(\theta)p_i, \qquad R_x(\theta) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & cos\theta & -sin\theta \\ 0 & sin\theta & cos\theta \end{bmatrix} \tag{5}$$

where $R_x(\theta)$ represents the rotation matrix around the x-axis with tilt angle θ. These transformations expose the model to perspective variations that commonly occur in real-world acquisition scenarios, such as differences in camera placement and user posture.

Parameter constraints and anatomical plausibility

To ensure semantic validity, all augmentation parameters are empirically constrained based on anatomical considerations. Excessive perturbations that could distort hand topology or alter the meaning of BISINDO alphabet gestures are explicitly avoided. This design choice aligns with prior findings that uncontrolled augmentation may degrade performance by introducing unrealistic samples [25], [27]. By balancing data diversity and anatomical plausibility, the proposed augmentation framework is tailored for low-resource BISINDO recognition scenarios.

Based on these considerations, all augmentation parameters are defined within fixed numerical ranges to ensure both consistency and reproducibility. For Gaussian Jitter, the noise standard deviation σ is sampled within the range [0.005, 0.015], which corresponds to an average bone-length distortion below 10%. Anisotropic scaling factors s_x, s_y, s_z are independently sampled from [0.8, 1.2], representing up to 20% variation in hand size while preserving anatomical plausibility. For Bone Length Scale, the scaling factor α is sampled within [0.9, 1.1], ensuring local bone adjustments remain subtle and biologically realistic. Depth variation Δz is sampled from a uniform distribution within ±25% of the observed depth range of detected keypoints, while tilt angles θ are sampled within [-15°, 15°] to simulate moderate viewpoint changes.

During training, augmentation is applied once to the training subset within each fold after K-Fold partitioning. The fold structure is fixed prior to augmentation, and only the training portion of each fold is expanded using oversampling. Each class is augmented until it reaches 300 samples per class, resulting in approximately 7,800 training samples per fold. Augmented samples are generated using randomly sampled

parameter values within the defined ranges, ensuring sufficient variability while maintaining consistent training data across epochs.

Model architecture and training strategy

Model architecture

The classification model in this study is implemented using a Multi-Layer Perceptron (MLP) architecture. The choice of MLP is motivated by the nature of the extracted features, which consist of low-dimensional, structured geometric vectors derived from normalized 3D hand keypoints [32]. In this setting, fully connected architectures are well suited to model non-linear relationships between joint coordinates without introducing the additional computational overhead associated with convolutional operations [8].

Previous studies on keypoint-based sign language recognition have shown that MLP-based classifiers can achieve competitive performance when the input features already encode meaningful geometric structure [33]. Following this line of work, the MLP is employed as a controlled and interpretable baseline to isolate the effect of the proposed augmentation strategy.

Network configuration

The network architecture consists of an input layer with 126 neurons corresponding to the flattened 3D keypoint features, followed by three hidden layers with 128, 64, and 32 neurons, respectively. Each hidden layer applies the Gaussian Error Linear Unit (GELU) activation function, which is selected for its smooth non-linearity and stable gradient propagation in deep networks.

To mitigate overfitting, dropout regularization is applied after each hidden layer with rates determined during hyperparameter optimization. The output layer employs a Softmax activation function to produce class probability distributions over the 26 BISINDO alphabet classes (A–Z).

Hyperparameter search

Hyperparameter optimization is conducted using a grid search strategy to systematically explore combinations of learning rate, batch size, and dropout configuration. The search space is defined based on preliminary experiments and prior studies on small-scale keypoint datasets, where excessive model capacity may lead to unstable training dynamics.

Rather than selecting the best configuration solely based on peak validation performance, this study prioritizes both accuracy and training stability. This consideration is particularly important given the limited size and homogeneous nature of the training data.

K-fold cross-validation

Model evaluation during training follows a K-Fold Cross-Validation scheme with $k = 4$. This choice reflects a balance between reliable performance estimation and data availability, as moderate fold sizes have been shown to provide more stable estimates on limited datasets [34].

In each fold, three partitions are used for training and one for validation. Data augmentation is applied exclusively to the training partitions within each fold, while the validation data remain unaltered. This protocol ensures that validation metrics reflect the model's generalization ability rather than exposure to augmented samples.

Stability-oriented model selection

To assess training stability across folds, the coefficient of variation (CV) of the validation F1-score is used as an auxiliary metric alongside average performance. A lower CV indicates more consistent model behavior across different data partitions, which is desirable in low-resource scenarios where model variance can be high [35]. The final model configuration is selected based on a joint criterion of high average validation performance and low CV, ensuring that the chosen model generalizes reliably across subsets of the training data.

Final training

After identifying the optimal hyperparameter configuration, the model is retrained using the entire internal training subset. The same augmentation strategy is applied to expand the dataset to the target distribution,

ensuring consistency with the training conditions used during cross-validation. This step allows the network to leverage all available training samples before final evaluation. Retraining on the full dataset is a standard practice to maximize model capacity once validation-based model selection has been completed [36].

Evaluation protocol

To ensure a clear and leakage-free evaluation protocol, the internal dataset is first divided into training (80%) and held-out test (20%) subsets using stratified sampling. The held-out test set is strictly reserved for final internal evaluation and is not involved in any training or model selection process.

The training subset is then used for model development through 4-fold cross-validation. In each fold, the data is partitioned into training and validation, where augmentation is applied only to the training portion after splitting. This ensures that no augmented samples appear in the validation data and prevents data leakage during model selection.

After determining the optimal hyperparameter configuration, the model is retrained using the entire training subset (80%). The final model is subsequently evaluated on the held-out internal test set, and further assessed on the external dataset for cross-domain evaluation. A summary of the data splitting and evaluation procedure is presented in Table 1. No samples from the held-out test set or external dataset are involved at any stage of model selection or training.

Table 1. Overview of data splitting, augmentation, and evaluation protocol

Stage	Data Portion	Description	Augmentation
Initial Split	80% Training / 20% Test	Stratified split applied before training	No
Cross-Validation	4-Fold (Training Set)	Model selection using training/validation folds	Training fold only
Augmentation (CV)	Training folds	Oversampled to 300 samples per class	Yes
Final Training	Full Training Set (augmented)	Model retrained with optimal hyperparameters	Yes
Internal Evaluation	Held-out Test Set (20%)	Evaluation under in-domain setting	No
External Evaluation	External Dataset (520)	Evaluation under cross-domain setting	No

Evaluation metrics

Model performance is evaluated quantitatively using a confusion matrix–based analysis. From the confusion matrix, four standard multi-class classification metrics are computed: Accuracy, Precision, Recall, and F1-Score [37]. These metrics are selected to provide a balanced assessment of overall correctness, class-wise prediction reliability, and sensitivity to false predictions.

Among these metrics, the F1-Score is emphasized as the primary indicator of performance, as it captures the harmonic balance between precision and recall. This is particularly relevant for BISINDO alphabet recognition, where certain classes may exhibit higher confusion due to geometric similarity between hand gestures, and relying solely on accuracy could mask class-specific errors.

These metrics are derived from the confusion matrix, which summarizes prediction outcomes into true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN). Accuracy measures the proportion of correctly classified samples among all predictions, while Precision indicates how many predicted instances of a class are actually correct. Recall measures the ability of the model to correctly identify all instances belonging to a class, and the F1-Score combines precision and recall through their harmonic mean to provide a balanced measure of classification performance. The mathematical formulations used to compute these metrics from the confusion matrix are presented in Equations (6) - (9).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (9)$$

Internal evaluation
Internal evaluation is conducted using the held-out test subset derived from the internal dataset. This evaluation reflects the model's ability to recognize BISINDO alphabet gestures under conditions that closely resemble the training distribution. While high internal performance indicates successful feature learning, it does not necessarily imply robustness to unseen variations in subject anatomy or acquisition devices.

Cross-domain evaluation
To rigorously assess generalization capability, this study employs a cross-domain evaluation protocol using an external dataset that differs from the training data in both subject identity and acquisition device. The trained models are evaluated directly on this dataset without any additional fine-tuning or adaptation.

This evaluation scenario simulates real-world deployment conditions, where sign language recognition systems are expected to operate on users and devices not encountered during training. Performance under this setting serves as a critical indicator of whether the learned representations capture subject-invariant geometric patterns rather than memorizing characteristics specific to the training data.

Comparative evaluation design
Both the baseline model and the augmented model are evaluated using identical evaluation protocols and datasets. This controlled comparison ensures that observed performance differences can be attributed directly to the proposed augmentation strategy rather than discrepancies in evaluation conditions.

In addition to quantitative metrics, confusion matrices are analyzed to identify recurring misclassification patterns, particularly among gesture classes with high geometric similarity. This qualitative analysis provides further insight into how augmentation influences the model's decision boundaries under cross-domain conditions.

RESULT AND DISCUSSION

Internal evaluation results
Internal evaluation was performed to measure in-domain recognition performance and to examine whether the introduction of realistic 3D keypoint augmentation negatively affects classification accuracy when training and testing data originate from the same distribution. In this stage, the baseline model trained without augmentation was compared with the proposed model trained using the augmented keypoint data on the internal test set.

As reported in Table 2, the baseline model achieved an F1-score of 97.07%, indicating strong recognition capability under single-subject conditions. When realistic 3D keypoint augmentation was applied, the model performance increased further, reaching an F1-score of 99.02%. Consistent improvements were also observed in accuracy, precision, and recall, suggesting that the augmentation strategy does not compromise in-domain performance.

Table 2. Internal evaluation performance of the baseline and augmented models on the internal test set

Model	Accuracy	Precision	Recall	F1-Score
Baseline	96.97%	97.69%	97.16%	97.07%
Proposed	98.99%	99.23%	99.04%	99.02%

Beyond aggregate metrics, qualitative inspection revealed that the augmented model produced fewer misclassifications among alphabet classes with similar hand configurations. This behavior indicates that the introduced geometric variations act as a regularization mechanism, encouraging the model to learn more stable decision boundaries even when evaluated on data from the same acquisition domain.

Importantly, these findings confirm that the proposed augmentation strategy enhances robustness without introducing a trade-off between generalization and internal accuracy. This result establishes a reliable basis for subsequent cross-domain evaluation, where the model is tested under more challenging conditions involving unseen subjects and devices.

Cross-domain evaluation results

Cross-domain evaluation was conducted to assess model robustness when exposed to data collected from unseen subjects and different acquisition devices. Unlike internal evaluation, this scenario represents realistic deployment conditions where variations in hand anatomy, gesture execution style, and camera characteristics are unavoidable. Both the baseline and augmented models were evaluated using the external dataset, which was excluded entirely from the training and validation processes.

The quantitative results summarized in Table 3 reveal a substantial performance gap between the two models. The baseline model suffered a sharp decline in recognition performance, achieving an F1-score of 67.20%. This degradation indicates strong overfitting to subject-specific and device-specific patterns learned during single-subject training. In contrast, the proposed model incorporating realistic 3D keypoint augmentation achieved a markedly higher F1-score of 89.55%, demonstrating significantly improved generalization capability.

Table 3. Cross-domain evaluation results comparing the baseline model and the proposed model

Model	Accuracy	Precision	Recall	F1-Score
Baseline	70.39%	69.18%	70.38%	67.2%
Proposed	89.62%	90.80%	89.62%	89.55%

Further insight is provided by the confusion matrix shown in Figure 5. The baseline model exhibits frequent misclassification among alphabet pairs with similar geometric structures, reflecting its limited ability to handle domain shifts. After augmentation, these error patterns were noticeably reduced, indicating that the model relies less on subject-specific cues and more on invariant geometric relationships.

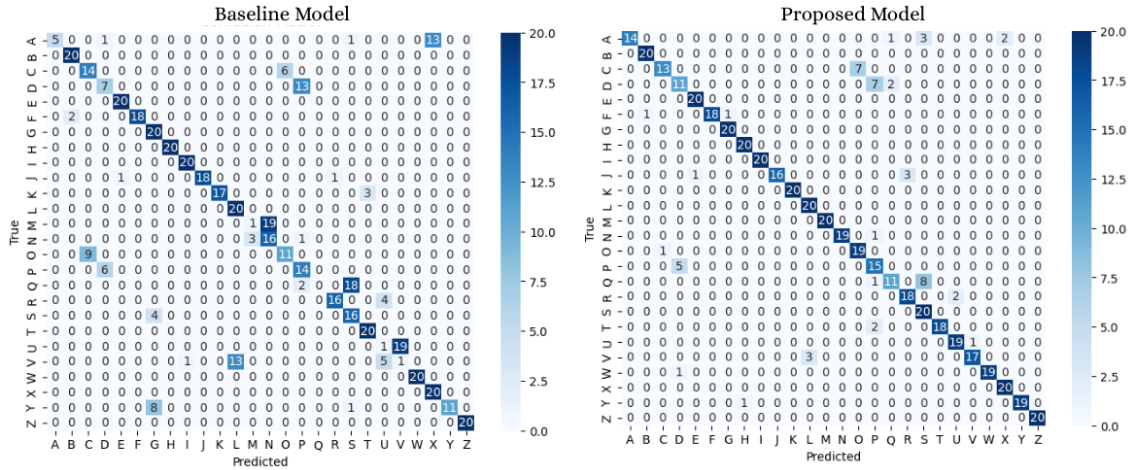

Figure 5. Confusion matrix of cross-domain BISINDO alphabet recognition

Overall, the contrast between internal and cross-domain performance highlights the limitations of training models solely on homogeneous datasets. While the baseline model performs well under controlled conditions, it fails to generalize to unseen environments. The augmented model, however, maintains a much smaller performance gap, suggesting that the introduced geometric variability effectively approximates real-world conditions during training.

Stability analysis across cross-validation folds

To evaluate the consistency of the observed performance gains, a stability analysis was conducted across the k-fold cross-validation process on the internal dataset. This analysis aims to determine whether the

improvements achieved through augmentation persist across different data partitions or arise from favorable data splits.

As presented in Table 4, the baseline model shows relatively high variability across folds, with a coefficient of variation (CV) of approximately 3.43%. This level of variability indicates sensitivity to changes in training data composition. In contrast, the augmented model demonstrates substantially more consistent behavior, achieving a CV of approximately 0.63%, which reflects stable performance across different folds.

Table 4. Stability analysis across k-fold cross-validation for the baseline and augmented models

Model	Ranks	Learning Rate	Batch Size	Dropout (L1, L2, L3)	F1-Score Mean	Coefficient of Variation
Baseline	1	0.001	64	0.5, 0.3, 0.1	94.3676%	3.4263%
	2	0.0005	32	0.5, 0.3, 0.1	88.737%	3.5148%
	3	0.001	32	0.6, 0.2, 0.1	88.891%	3.6852%
Proposed	1	0.001	64	0.5, 0.3, 0.1	98.0952%	0.6334%
	2	0.001	32	0.6, 0.2, 0.1	98.0502%	0.6837%
	3	0.0005	64	0.5, 0.3, 0.1	97.4451%	0.7001%

Training dynamics further support this observation. As illustrated in Figure 6, the augmented model converges significantly faster and exhibits smoother loss curves compared to the baseline model. While the baseline model requires up to 108 epochs to stabilize, the augmented model reaches convergence within approximately 20 epochs, indicating more efficient optimization and reduced susceptibility to overfitting.

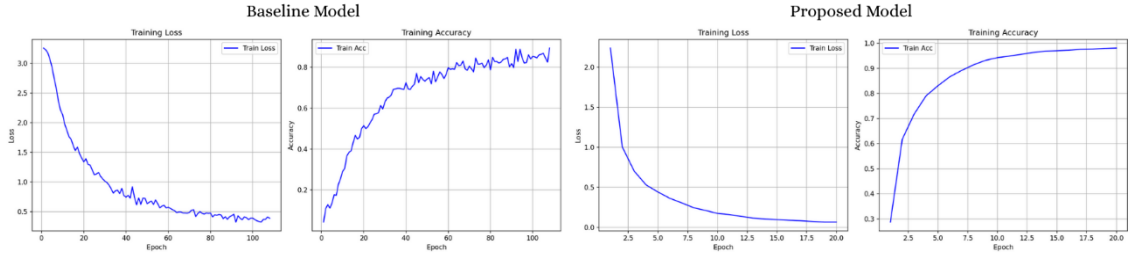

Figure 6. Training curves of the baseline and augmented models

These results suggest that realistic 3D keypoint augmentation not only improves recognition accuracy but also contributes to more stable training behavior. Exposure to controlled geometric variability encourages the model to learn robust representations that generalize consistently across different training subsets.

Effects of realistic 3D keypoint augmentation on model generalization

The observed improvements in cross-domain performance can be attributed to changes in how the model encodes geometric relationships among hand keypoints. Rather than depending on precise joint coordinates, training with augmented data encourages the model to emphasize relative spatial configurations that remain consistent across subjects and recording conditions. This shift away from memorization is particularly important in single-subject training scenarios.

Each augmentation component contributes to generalization through a distinct mechanism. Gaussian Jitter increases tolerance to sensor noise and minor execution variability, preventing excessive sensitivity to exact joint positions. Anisotropic Scale and Bone Length Scale expose the model to differences in hand proportions and finger lengths that naturally occur across individuals. Depth & Tilt Jitter further enhances robustness by simulating variations in camera distance and orientation. Together, these complementary transformations produce a more expressive and resilient feature space than any individual augmentation applied in isolation.

These findings align with prior research in skeleton-based action recognition and hand pose analysis, which reports that feature-level geometric augmentation is more effective than image-level manipulation when semantic structure must be preserved. By constraining augmentation parameters within anatomically plausible ranges, the proposed approach avoids generating unrealistic samples that could degrade

performance. Consequently, the performance gains observed in this study arise not merely from increased data diversity, but from the introduction of variability that closely reflects real-world acquisition conditions.

Limitations and practical considerations

Despite the demonstrated effectiveness of the proposed augmentation framework, several limitations remain. First, the study focuses exclusively on static BISINDO alphabet recognition and does not consider temporal dynamics inherent in continuous sign language. As a result, the current findings may not directly generalize to sequence-based recognition tasks that require modeling motion transitions over time. Second, while the augmentation strategy improves robustness to subject and device variation, it does not explicitly address challenging conditions such as occlusion, extreme hand poses, or highly cluttered backgrounds.

From a practical standpoint, the proposed approach is well suited for low-resource sign language recognition scenarios. Because augmentation is performed at the feature level, it introduces minimal computational overhead and can be integrated into existing keypoint-based pipelines without modifying upstream pose estimation components. This makes the method particularly suitable for deployment on mobile and edge devices. Moreover, although this study focuses on BISINDO, the augmentation framework is general and can be adapted to other sign languages or gesture recognition tasks that rely on skeletal representations.

CONCLUSION

This study demonstrates that realistic 3D keypoint augmentation is an effective strategy for improving the generalization capability of BISINDO alphabet recognition models trained under limited and homogeneous data conditions. By introducing controlled geometric variations at the feature level, the proposed approach mitigates overfitting without increasing model complexity or requiring additional data collection. The experimental results indicate that the proposed model achieves strong performance, reaching an F1-score of 99.02% in internal evaluation and 89.55% under cross-domain testing, reflecting both high in-domain accuracy and improved robustness to unseen subjects and different acquisition devices.

Beyond performance gains, this work highlights the importance of data-centric design in keypoint-based sign language recognition, where carefully constrained augmentation can encourage models to learn invariant geometric representations rather than subject-specific patterns. The findings also suggest that introducing realistic variability during training plays a key role in reducing performance gaps between controlled and real-world conditions.

Despite these contributions, several limitations remain. This study focuses only on static alphabet recognition and does not capture temporal dynamics in continuous sign language. In addition, while geometric augmentation improves robustness, it cannot fully represent all variations in real hand anatomy and complex acquisition scenarios. Based on these limitations, future work is directed toward expanding dataset diversity with more real subjects and adapting the proposed augmentation strategy to dynamic gesture recognition using sequence-based models such as LSTM or Transformer. These directions are expected to further improve the applicability of the proposed approach in real-world sign language recognition systems.

REFERENCES

[1] S. A. Satillah, K. Khotimah, N. Nisai Muslihah, W. Mirip, Y. Dwi Suryani, and L. Dwi Ayu Pagarwati, “Ragam Bahasa Anak Tunarungu Dengan SIBI Di SLB N Ogan Ilir,” vol. 07, no. 01, pp. 31–43, 2024, [Online]. Available: https://jurnal.uai.ac.id/index.php/AUDHI

[2] A. Fathiray, J. Maulindar, and W. Lestari, “Pengembangan Sistem Penerjemah Kalimat Bahasa Isyarat Bisindo To Text Dengan Kinect Real Time,” *Infotek: Jurnal Informatika dan Teknologi*, vol. 8, pp. 1–12, Jan. 2025, doi: 10.29408/jit.v8i1.26116.

[3] H. M. Monisha, “Sign Language Detection and Classification using Hand Tracking and Deep Learning in Real-Time,” pp. 2356–2395, Nov. 2023.

[4] N. Nugroho and M. Putra, “LEVERAGING DEEP LEARNING APPROACH FOR ACCURATE ALPHABET RECOGNITION THROUGH HAND GESTURES IN SIGN LANGUAGE,” *Jurnal Teknik Informatika (Jutif)*, vol. 6, pp. 259–268, Feb. 2025, doi: 10.52436/1.jutif.2025.6.1.3912.

[5] M. R. Ningsih, Y. J. Nurriski, F. A. Z. Sanjani, M. F. Al Hakim, J. Unjung, and M. A. Muslim, "Sign Language Detection System Using YOLOv5 Algorithm to Promote Communication Equality People with Disabilities," *Scientific Journal of Informatics*, vol. 11, no. 2, pp. 549–558, Jun. 2024, doi: 10.15294/sji.v11i2.6007.

[6] J. A. Rodríguez-Rodríguez, E. López-Rubio, J. A. Ángel-Ruiz, and M. A. Molina-Cabello, "The Impact of Noise and Brightness on Object Detection Methods," *Sensors*, vol. 24, no. 3, Feb. 2024, doi: 10.3390/s24030821.

[7] A. A. Volkov *et al.*, "A Review of Neural Network-Based Image Noise Processing Methods," *Sensors*, vol. 25, no. 19, p. 6088, Oct. 2025, doi: 10.3390/s25196088.

[8] A. B. Wardana, J. Armyanto, and E. Daniati, "Perancangan Algoritma SVM untuk Pengembangan Model Pendeteksi Bahasa Isyarat Berbasis Landmark," Online, 2025.

[9] D. K. Gonfa, M. Meshesha, D. W. Girmaw, and K. W/Maryam, "A deep learning framework for Ethiopian sign language recognition using skeleton-based representation," *Sci. Rep.*, vol. 15, no. 1, p. 36181, 2025, doi: 10.1038/s41598-025-19937-0.

[10] R. Yayla, H. Üçgün, and M. Abbas, "Pose-Based Static Sign Language Recognition with Deep Learning for Turkish, Arabic, and American Sign Languages," *Sensors*, vol. 26, no. 2, 2026, doi: 10.3390/s26020524.

[11] A. D. Goenawan and S. Hartati, "The Comparison of K-Nearest Neighbors and Random Forest Algorithm to Recognize Indonesian Sign Language in a Real-Time," *Scientific Journal of Informatics*, vol. 11, no. 1, pp. 237–244, Feb. 2024, doi: 10.15294/sji.v11i1.48475.

[12] F. Zhang *et al.*, "MediaPipe Hands: On-device Real-time Hand Tracking," Jun. 2020, [Online]. Available: http://arxiv.org/abs/2006.10214

[13] R. Wiliam, C. Lufian, Meiliana, and A. Y. Zakiyyah, "Recognitions of Bahasa Isyarat Indonesia (BISINDO) Alphabets Using SVM and Mediapipe," in *2024 6th International Conference on Cybernetics and Intelligent System (ICORIS)*, IEEE, Nov. 2024, pp. 1–5. doi: 10.1109/ICORIS63540.2024.10903898.

[14] 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.

[15] M. Yousef and J. Allmer, "Deep learning in bioinformatics," *Turkish Journal of Biology*, vol. 47, no. 6, pp. 366–382, 2023, doi: 10.55730/1300-0152.2671.

[16] A. Muttaqin, A. Luthfiarta, A. Nugraha, and P. Salsabila, "Single-Image Face Recognition For Student Identification Using Facenet512 And Yolov8 In Academic Environtment With Limited Dataset," *Jurnal Teknik Informatika (Jutif)*, vol. 6, pp. 3018–3032, Oct. 2025, doi: 10.52436/1.jutif.2025.6.5.3908.

[17] C. Rohlfs, "Generalization in neural networks: A broad survey," *Neurocomputing*, vol. 611, p. 128701, 2025, doi: https://doi.org/10.1016/j.neucom.2024.128701.

[18] D. K. Ahwani, S. N. Budiman, and M. F. Rahmat, "Penerapan Mediapipe dalam Pengenalan Bisindo Berbasis Deep Learning dan Computer Vision (Studi Kasus: SLB-C B Yayasan Pendidikan Luar Biasa (YPLB) Blitar)," *INNOVATIVE: Journal Of Social Science Research*, vol. 4, pp. 3608–3628, 2024.

[19] A. Mumuni and F. Mumuni, "Data augmentation: A comprehensive survey of modern approaches," *Array*, vol. 16, p. 100258, 2022, doi: https://doi.org/10.1016/j.array.2022.100258.

[20] R. Satila Passa, S. Nurmaini, and D. P. Rini, "YOLOv8 Based on Data Augmentation for MRI Brain Tumor Detection," *Scientific Journal of Informatics*, vol. 10, no. 3, p. 363, 2023, doi: 10.15294/sji.v10i3.45361.

[21] C. Xin, S. Kim, and K. Park, *A Comparison of Machine Learning Models with Data Augmentation Techniques for Skeleton-based Human Action Recognition*. 2023. doi: 10.1145/3584371.3612999.

[22] Y. Peng, P. Jiao, H. Zou, Y. Min, and X. Chen, "Synthetic View Augmentation for Sign Language Recognition," in *Companion Proceedings of the ACM on Web Conference 2025*, in WWW '25. New York, NY, USA: Association for Computing Machinery, 2025, pp. 2448–2452. doi: 10.1145/3701716.3717520.

[23] C. Xin, S. Kim, Y. Cho, and K. S. Park, "Enhancing Human Action Recognition with 3D Skeleton Data: A Comprehensive Study of Deep Learning and Data Augmentation," *Electronics (Switzerland)*, vol. 13, no. 4, Feb. 2024, doi: 10.3390/electronics13040747.

[24] K. Seo, H. Cho, D. Choi, and J.-D. Park, "Implicit Semantic Data Augmentation for Hand Pose Estimation," *IEEE Access*, vol. 10, pp. 84680–84688, 2022, doi: 10.1109/ACCESS.2022.3197749.

[25] Y. Nakamura and L. Jing, "Skeleton-Based Data Augmentation for Sign Language Recognition Using Adversarial Learning," *IEEE Access*, vol. 13, pp. 15290–15300, 2025, doi: 10.1109/ACCESS.2024.3481254.

[26] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *J. Big Data*, vol. 6, no. 1, p. 60, 2019, doi: 10.1186/s40537-019-0197-0.

[27] M. H. Putri, U. Zaky, and B. A. Prabawa, "Optimizing Data Augmentation Parameters in YOLOv11 for Enhanced Rip Current Detection on Small Datasets from Depok-Parangtritis Coastline," *Jurnal Teknik Informatika (JUTIF)*, vol. 6, no. 5, 2025, doi: 10.52436/1.jutif.6.5.5352.

[28] K. Roh, H. Lee, E. J. Hwang, S. Cho, and J. C. Park, "Preprocessing Mediapipe Keypoints with Keypoint Reconstruction and Anchors for Isolated Sign Language Recognition," in *Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*, E. Efthimiou, S.-E. Fotinea, T. Hanke, J. A. Hochgesang, J. Mesch, and M. Schulder, Eds., Torino, Italia: ELRA and ICCL, May 2024, pp. 323–334. [Online]. Available: https://aclanthology.org/2024.signlang-1.36/

[29] M. Schlegel and K.-U. Sattler, "Management of Machine Learning Lifecycle Artifacts: A Survey," *SIGMOD Rec.*, vol. 51, no. 4, pp. 18–35, Jan. 2023, doi: 10.1145/3582302.3582306.

[30] D. Joan, V. Vincent, K. J. Daniel, S. Achmad, and R. Sutoyo, "BISINDO Hand-Sign Detection Using Transfer Learning," in *2023 IEEE 8th International Conference on Recent Advances and Innovations in Engineering (ICRAIE)*, 2023, pp. 1–7. doi: 10.1109/ICRAIE59459.2023.10468194.

[31] U. Iqbal, P. Molchanov, and J. Kautz, *Weakly-Supervised 3D Human Pose Learning via Multi-View Images in the Wild*. 2020. doi: 10.1109/CVPR42600.2020.00529.

[32] D. Sinaga, C. Jatmoko, S. Suprayogi, and N. Hedriyanto, "Multi-Layer Convolutional Neural Networks for Batik Image Classification," *Scientific Journal of Informatics*, vol. 11, no. 2, pp. 477–484, May 2024, doi: 10.15294/sji.v11i2.3309.

[33] N. H. Amir, C. K. Dewa, and A. Luthfi, "Refining the Performance of Neural Networks with Simple Architectures for Indonesian Sign Language System (SIBI) Letter Recognition Using Keypoint Detection," *ILKOM Jurnal Ilmiah*, vol. 17, no. 1, pp. 64–73, Apr. 2025, doi: 10.33096/ilkom.v17i1.2522.64-73.

[34] D. Radočaj, M. Jurišić, I. Plaščak, and L. Galić, "Randomness in Data Partitioning and Its Impact on Digital Soil Mapping Accuracy: A Comparison of Cross-Validation and Split-Sample Approaches," *Agronomy*, vol. 15, no. 11, Nov. 2025, doi: 10.3390/agronomy15112495.

[35] R. Walpole, S. Myers, R. Myers, and K. Ye, *Probability & Statistics for Engineers & Scientists : Global Edition*. Pearson Deutschland, 2024. doi: DOI.

[36] S. Raschka, *Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning*. 2018. doi: 10.48550/arXiv.1811.12808.

[37] R. A. Rumy, A. Al Maruf, and Z. Aung, "Multi-label emotion and sentiment detection from Bengali text using hybrid deep learning model," *International Journal of Cognitive Computing in Engineering*, vol. 7, pp. 213–234, 2026, doi: https://doi.org/10.1016/j.ijcce.2025.10.005.