Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871

Vol. 7, No. 2, April 2026, Page. 823-842
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5210

Performance Analysis of Traditional Machine Learning Classifiers on LSTM-Extracted Features for Indonesian Sign Language System Recognition

Patricia Ho*[1], Handri Santoso[2]

[1,2]Informatics, Universitas Pradita, Indonesia

Email: [1]patricia.ho@student.pradita.ac.id

Received : Jul 31, 2025; Revised : Aug 14, 2025; Accepted : Aug 15, 2025; Published : Apr 15, 2026

Abstract

Recognizing affix gestures in the Indonesian Sign Language System (SIBI) remains challenging due to subtle visual differences in hand shape and movement, often resulting in lower classification accuracy compared to other categories. This study aims to evaluate whether lightweight traditional and hybrid classifiers can provide competitive performance to deep learning models for SIBI recognition. Using a dataset of 21,351 gesture videos covering four categories (Affix, Alphabet, Number, and Word), features were extracted from MediaPipe keypoints and processed as frozen LSTM embeddings. Six classifiers (Random Forest, K-Nearest Neighbors, Naïve Bayes, Multilayer Perceptron, Support Vector Machine, and Hidden Markov Model) were evaluated with 5-fold stratified cross-validation using accuracy, precision, recall, and F1-score, with statistical significance tested through Friedman and Nemenyi analyses. Results show that MLP and RF achieved high performance in Alphabet, Number, and Word categories (above 96 percent accuracy), while Affix remained the most difficult, with MLP reaching 81.17 percent, outperforming the 68.17 percent from a prior BiLSTM model. This study provides a benchmark for hybrid model implementation in sign language recognition, showing that while traditional classifiers on deep features are effective and computationally lighter for general gestures, deep architectures remain superior for capturing the fine-grained temporal nuances critical for complex categories like affixes.

Keywords : *Deep Learning, Hand Gesture Recognition, Indonesian Sign Language System (SIBI), Machine Learning, MediaPipe.*

1. INTRODUCTION

Building upon the success of our previous study, "LSTM-Based Hand Gesture Recognition for Indonesian Sign Language System (SIBI) on Affix, Alphabet, Number, and Word" published in June 2025, this research continues to explore recognition performance across individual gesture categories. The earlier work showed that a BiLSTM-based framework could reliably recognize most SIBI gestures, achieving over 91% accuracy for the alphabet, number, and word categories [1].

However, the affix category presented significant challenges, with accuracy dropping to around 68%. This lower performance was largely due to the nature of affix gestures, which often share highly similar hand orientations and motions, differing only in subtle aspects such as thumb placement, slight finger rotations, or minimal bending. Similar fine-grained recognition challenges have been reported in fingerspelling tasks [2]. These fine-grained differences not only make it difficult for automated models to distinguish between signs like "se," "me," and "ter," but also pose challenges even for human observers [1].

Efforts to improve recognition accuracy for challenging gesture categories like affixes have explored various directions. Optical flow-based temporal segmentation has been used to separate continuous SIBI sentences into isolated root and affix signs, facilitating downstream recognition. Conceptual frameworks

have also emphasized the need for total communication, integrating manual cues (hand shapes and motions) with non-manual cues (facial expressions and posture), to produce semantically richer translations. While deep architectures such as CNN-LSTM and Transformer-based models have shown promise in capturing spatial-temporal dynamics, their high computational costs and data requirements make them less ideal for lightweight, real-time applications [3], [4].

Deep learning models like BiLSTM are powerful at learning sequential and temporal patterns. However, they are not always enough for gestures that require very fine visual discrimination. Research on sign language kinematics has shown that each signer tends to have a unique style of motion, characterized by distinct velocity patterns and joint movement relationships, regardless of the actual sign being performed [5]. When deep models are trained on datasets that are small or unbalanced, they are more prone to overfitting [6], memorizing training samples rather than learning generalized patterns, a limitation widely reported in small-data scenarios and often addressed with techniques such as transfer learning, data augmentation, and synthetic data generation [7]. In addition, in imbalanced datasets, the model can become biased toward majority classes, causing poor recognition of minority classes and amplifying misclassification for subtle, fine-grained categories [8]. As a result, instead of learning the true, subtle differences between gesture classes, the model might rely on the way a specific person moves, which can hurt accuracy for categories with only slight visual differences, such as affixes.

To address problems like these, recent studies have explored hybrid models that combine deep learning and traditional machine learning methods. These hybrid systems take advantage of what each approach does best. For example, an LSTM-SVM hybrid model has been used in failure prediction tasks, where the LSTM handles sequential data and the SVM performs the final classification [9], [10]. In other research, CNN-LSTM models have improved sign language recognition by using CNN layers to extract spatial features from video frames while LSTM layers learn the temporal patterns over time [11]. Hybrid models have also shown strong results outside of gesture recognition. In one industrial application, a system that combined LSTM with a Random Forest classifier, along with feature selection using Grey Wolf Optimization (GWO), reached 98.97 percent accuracy. This performance was higher than using LSTM alone (93.56 percent) or Random Forest alone (98.44 percent). The hybrid approach was especially effective at distinguishing between classes that were very similar to each other, because the LSTM could capture complex temporal patterns and the Random Forest could make more robust classification decisions [12]. In hyperspectral image classification, hybrid frameworks integrating convolutional neural networks with traditional classifiers such as support vector machines have been shown to significantly improve accuracy over either method alone, highlighting the adaptability of this approach across domains [13]. In a related example, hybrid architectures that pair LSTM with non-linear tree-based models have also demonstrated strong performance in complex classification tasks. An LSTM-Decision Tree framework designed to handle both sequential and categorical features achieved higher accuracy, precision, recall, and F1-score than either model alone when tested on a multi-year university dataset [14].

Despite these successes in other domains, hybrid modeling has not yet been widely applied to SIBI gesture recognition, especially for addressing the unique challenges of affix gestures. Most existing studies either rely completely on deep learning or only make a limited comparison with traditional classifiers. Few have tested hybrid frameworks in this specific context, even though the ability to capture subtle hand shapes and motion cues is critical for improving accuracy [15]. Prior Indonesian sign-language studies confirm the value of temporal modeling, LSTM for SIBI and real-time Indonesian sign language, and even sentence generation pipelines [16], [17], [18], while computer-vision hand gesture recognition shows that deep features with classical heads such as CNN embeddings with SVM can be highly effective, and pure LSTM streams also remain strong [19], [20], [21]. Cross-domain evidence further supports hybridization, such as compact LSTM–SVM models in cardiovascular screening [22] and LSTM features with Random Forests

Vol. 7, No. 2, April 2026, Page. 823-842
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5210

for flood prediction both report high accuracy with lightweight decision heads [23]. Classical baselines like Naïve Bayes have been used for hand gesture recognition, and broader sequence studies indicate BiLSTM tends to surpass Naïve Bayes standalone [24], [25].

Motivated by these gaps and cross-domain evidence, a hybrid SIBI pipeline is adopted. In this approach, the LSTM model is not used as the final classifier but instead acts as a feature extractor. It learns the temporal dynamics of hand movements across the video sequences and outputs representations of those patterns. These representations are then passed to traditional machine learning classifiers for the final decision-making step. To test the effectiveness of this approach, we compare several hybrid pipelines against a range of baseline classifiers, including K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Multilayer Perceptron (MLP), Random Forest (RF), Naive Bayes (NB), and Hidden Markov Models (HMM). These comparisons use multiple performance measures, such as accuracy, F1-score, and processing efficiency, to evaluate not just raw classification performance but also how suitable each method is for SIBI gesture recognition.

2. METHOD

The study starts by cleaning the raw SIBI video corpus until every clip is a neat, fixed-length gesture sequence. The original dataset contains four gesture groups, which are affix, alphabet, number, and sentence. First, any introductions and outros are removed, and the sentence clips are manually split into single-word videos. Every clip is then resampled to exactly 30 frames so that all sequences share the same temporal length. MediaPipe detects 21 hand key-points per frame, producing x and y coordinates for both hands. Missing points are filled by linear interpolation, while a binary mask records where interpolation occurred. After this procedure, the refined dataset holds 18 affix classes, 26 alphabet letters, 35 numbers, and 29 common words.

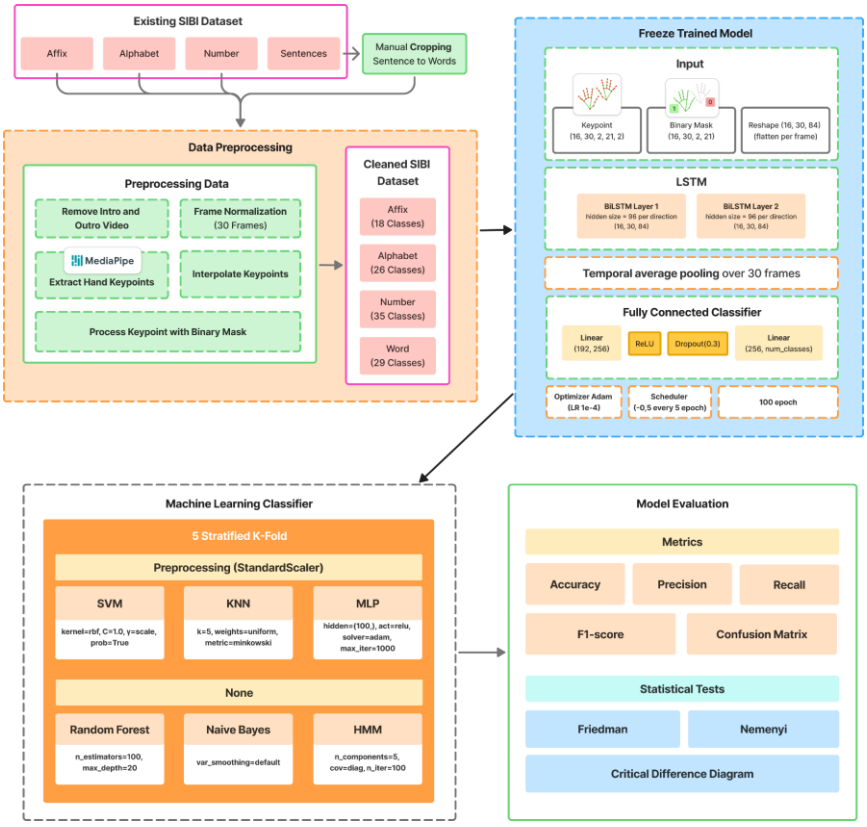


Figure 1. Detail Implementation Flow for Experiment

Vol. 7, No. 2, April 2026, Page. 823-842
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5210

Next, the key-point sequences pass through a two-layer bidirectional LSTM. Each direction in the LSTM has 96 hidden units, giving a combined output size of 192. The network is trained once and then frozen so that its weights do not change further. For every video, the hidden states from all 30 frames are averaged, resulting in 1 compact 192 dimensional feature vector that represents the entire gesture. Freezing the network ensures that later performance differences come only from the classifiers rather than further tuning of the deep model.

These frozen LSTM embeddings are provided to a set of lightweight classical machine-learning algorithms. Six classifiers are explored: Support Vector Machine (SVM), K-Nearest Neighbour (KNN), Random Forest (RF), Multi Layer Perceptron (MLP), Naïve Bayes (NB), and Hidden Markov Model (HMM). For SVM, KNN, and MLP, the input features are standardised using StandardScaler to ensure all features have zero mean and unit variance, as these algorithms are sensitive to feature magnitude. Random Forest, Naïve Bayes, and HMM do not require feature scaling because they are scale-invariant or operate on probability distributions and sequential likelihoods. Training and testing use 5-fold stratified cross-validation, which keeps the class proportions identical in every fold [26].

Finally, each model is evaluated with accuracy, precision, recall, F1-score, and confusion matrices. Mean and standard-deviation values across folds highlight robustness. Because multiple algorithms share the same data splits, a Friedman test first checks if their average ranks differ significantly. When the test is significant, a Nemenyi post-hoc analysis is applied, and the results are visualised through a Critical Difference (CD) diagram to identify exactly which classifier pairs perform differently.

2.1. Dataset and Data Preprocessing

The dataset used in this study is identical to that of our prior work, which focused on BiLSTM-based recognition of SIBI hand gestures across four categories: affix (18 classes), alphabet (26 classes), number (35 classes), and word (29 classes made from 10 sentence-derived classes). The recordings were collected from 20 to 22 subjects, each performing ten per gesture for affix, alphabet, and number; five repetitions per gesture with only neutral-emotion samples retained for analysis [27]. Each recording was manually trimmed to isolate the target gesture, excluding introductory and concluding motions to ensure consistency.

From each processed video, 30 evenly spaced RGB frames were sampled per gesture. All frames were resized to 224×224 pixels. Hand landmarks were extracted using MediaPipe, resulting in a (30, 2, 21, 2) array of keypoints per sample, along with a corresponding binary mask (30, 2, 21) to flag missing detections. Any NaN values in the keypoints were replaced with zeros prior to flattening. Both keypoint features and pixel values were standardized using statistics calculated exclusively from the training set to avoid information leakage, ensuring a fair evaluation.

2.2. LSTM Feature Extraction

For feature extraction, we used the same BiLSTM architecture from our earlier study, but with its weights frozen after training. Each sample's keypoint sequence was passed through the BiLSTM, and the final hidden states from the last layer were extracted as feature embeddings. These embeddings represent the temporal and spatial dynamics of each gesture sequence and serve as the sole input for the traditional machine learning classifiers tested in this work. No additional fine-tuning of the BiLSTM was performed, ensuring that all classifiers were evaluated on the same fixed feature set.

2.3. Traditional Classifiers and Hyperparameters

2.3.1. K-Nearest Neighbors (KNN)

KNN is a simple, instance-based supervised learning algorithm that classifies new gestures by comparing them to the closest labeled samples in the training set [28]. Using a distance metric (Euclidean

Vol. 7, No. 2, April 2026, Page. 823-842
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5210

in this study) and a fixed number of neighbors (k=5), KNN assigns the class most common among those neighbors. Its advantages include minimal hyperparameter tuning and no assumptions about data distribution, making it suitable for quick prototyping. However, it is computationally expensive for large datasets because distances must be recalculated for each prediction, and it performs poorly on high-dimensional data due to the "curse of dimensionality" [29].

2.3.2. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised learning algorithm that seeks to determine an optimal separating hyperplane by maximizing the margin between gesture classes in a high-dimensional feature space. According to the mathematical formulation used in the referenced study, this is achieved by minimizing the norm of the weight vector while satisfying the constraint that correctly classifies all training samples, effectively leading to a robust and generalizable decision boundary [30]. Using the Radial Basis Function (RBF) kernel (with C=1.0 and gamma set to 'scale' in this study), SVM can effectively handle non-linear and high-dimensional data, making it suitable for complex hand gesture recognition. Its advantages include strong generalization and ability to work with structured or semi-structured datasets, but it can be computationally intensive for large datasets and is sensitive to kernel selection [29].

Studies in hand gesture recognition have shown that SVM achieves strong generalization even with diverse user data and lighting conditions by leveraging preprocessed features such as HOG (Histogram of Oriented Gradients), Gaussian-thresholded silhouettes, and convex hull points. In comparative evaluations, different kernels achieved classification accuracies ranging from 24% to 90% [31].

2.3.3. Multilayer Perceptron (MLP)

The Multilayer Perceptron (MLP) is a feedforward deep neural network composed of an input layer, one or more hidden layers, and an output layer, where each neuron is fully connected to the next layer and trained using supervised learning through backpropagation [32], [33]. In this study, we configure the MLP with a single hidden layer of 100 neurons, using the Rectified Linear Unit (ReLU) activation to enable sparse and efficient gradient flow, and optimize the weights using the Adam optimizer. The maximum iteration count was 1000 to ensure convergence during training.

The design of this MLP is inspired by findings in recent deep learning research for gesture recognition, where lightweight fully connected layers are effectively used for late fusion of features from multiple modalities and temporal scales [34]. Additionally, a literature review in 2022 further supports the reliability of MLP as a classification algorithm, reporting an average accuracy of 91.98% across 30 studies, with the highest reaching 100% and the lowest 62.89% demonstrating MLP's strong applicability in both prediction and classification problem [33]. While complex convolutional and multi-scale networks can capture richer patterns, the MLP remains advantageous as a computationally efficient and generalizable alternative, particularly for tasks where feature extraction (from LSTM or keypoints) already captures most spatial and temporal context.

2.3.4. Random Forest (RF)

Random Forest (RF) is an ensemble learning method that builds multiple decision trees and combines their predictions via majority voting [35]. Each tree is trained on a random subset of data and features, which reduces overfitting and increases generalization. RF is particularly useful for gesture recognition because it handles noisy and missing data well and can model non-linear relationships. However, its prediction speed is slower due to the need to aggregate many trees, and the model behaves as a "black box," making it harder to interpret [29].

Unlike deep models, Random Forests have relatively low computational costs, especially when configured with moderate tree depth and estimator counts, making them viable for real-time system [36]. Studies have demonstrated that Random Forest-based gesture recognition can achieve competitive accuracy (93.07% on alphabet-based dataset) while training faster and requiring fewer resources than CNN-based models [37].

2.3.5. Naive Bayes (NB)

Naïve Bayes is a probabilistic classifier that uses Bayes' theorem under the assumption that all features contribute independently to the probability of a class [38]. For hand gesture recognition, it estimates the likelihood that a gesture belongs to a specific class based on features such as pixel intensity, edge orientation, or shape descriptors, multiplying the conditional probabilities of each feature and combining them with the prior probability of the class [29]. This simplicity makes NB computationally lightweight and highly efficient for real-time processing, as it requires minimal training and can scale well to large datasets.

Recent studies highlight NB's utility for early gesture recognition, where predictions must be made before the full gesture sequence is observed. Using a bag-of-features representation, NB can incrementally update its predictions as more frames are processed, leveraging its cumulative evidence property to classify gestures with only partial information. Experiments have shown that this approach can outperform more complex methods (max-margin SVMs) in early classification tasks while maintaining low computational cost [39].

However, NB's reliance on the independence assumption can limit its accuracy for gestures with highly correlated spatial-temporal features. Despite this, its speed, scalability, and ability to handle multi-class predictions make it a valuable baseline for hybrid frameworks, where NB can serve as a fast classifier for LSTM-derived features.

2.3.6. Hidden Markov Model (HMM)

Hidden Markov Models (HMMs) are probabilistic sequence models designed to capture the temporal dynamics of hand gestures by modeling them as transitions between hidden states [40]. In gesture recognition, each gesture class is represented as an HMM, where observed features (Hu moment invariants, centroid shifts, and area changes) are mapped to state emission probabilities. These states capture how gestures evolve over time, allowing the model to handle both temporal segmentation and classification simultaneously.

In this study, each HMM is configured with n_components=5 (the number of hidden states per gesture), using a covariance_type='diag' to assume diagonal covariance matrices for computational efficiency, and trained with up to n_iter=100 iterations for the Baum–Welch algorithm to ensure convergence. During inference, a normalized Viterbi algorithm is applied to stabilize state likelihoods across time and detect gesture occurrences by identifying peaks in model score [41].

2.4. Evaluation Protocol

To ensure fairness across all classifiers, we adopted 5-fold stratified cross-validation, which divides the dataset into five equal parts while preserving the class distribution in each fold. Each model is trained on four folds and validated on the remaining fold, rotating so that every fold is used for validation once. This strategy minimizes data imbalance across splits and allows direct comparison with our prior BiLSTM study.

In this study, model performance is evaluated using metrics derived from the confusion matrix, which summarizes classification outcomes across predicted and actual labels. For a binary classification case, the

Vol. 7, No. 2, April 2026, Page. 823-842
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5210

confusion matrix consists of four primary components, which are True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN) [42], representing correct positive detections, incorrect positive predictions, missed positives, and correct negative detections, respectively.

The Accuracy metric measures the proportion of correctly classified samples across all classes [43], defined as:

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

Aside from Accuracy, Precision and Recall are also used. Precision quantifies the proportion of correctly identified positive samples among all positive predictions [44], [45], expressed as:

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

In contrast, Recall, also referred to as Sensitivity or True Positive Rate (TPR), represents the proportion of actual positive samples correctly identified, It is calculated by dividing the number of true positives by the total number of actual positive instances (the sum of true positives and false negatives) [45].

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

To provide a single representative score that balances Precision and Recall, the F1-score is computed as the harmonic mean of the two metrics. A simple average of precision and recall can be misleading. For example, a model with a precision of 1.0 (100%) but a recall of 0.1 (10%) would have a simple average of 0.55. This score masks the fact that the model is failing to find 90% of the positive cases. The harmonic mean, however, heavily penalizes models where one of the two metrics is very low. In the same example, the F1 Score would be only 0.18. To achieve a high F1 Score, a model must perform reasonably well on *both* precision and recall, making it a truly "balanced" measure of performance [46].

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

To assess whether the performance differences among classifiers were statistically significant, we used the Friedman test followed by the Nemenyi post-hoc test, following the guidelines proposed by Demšar [47]. The Friedman test ranks each classifier for every fold based on its performance, then evaluates whether the average ranks across classifiers differ more than would be expected by chance. This test is non-parametric and does not assume normal distribution, making it suitable for comparing machine learning models evaluated across multiple datasets or folds.

The Friedman test statistic is computed as:

$$x_F{}^2 = \frac{12N}{k(k+1)} \times \left(\sum_{j=1}^{k} R_j{}^2 - \frac{k(k+1)^2}{4}\right) \quad (5)$$

In this equation, N represents the number of datasets or cross-validation folds used in the evaluation, k denotes the number of classifiers being compared, and R_j is the average rank of the j-th classifier across all folds. The test examines whether these average ranks differ significantly from each other under the null hypothesis that all classifiers perform equally.

If the Friedman test shows a significant result ($p < 0.05$), we apply the Nemenyi post-hoc test to determine which pairs of classifiers are significantly different from each other. The test calculates a critical difference (CD), and if the average rank difference between two classifiers exceeds this value, their

Vol. 7, No. 2, April 2026, Page. 823-842
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5210

performance is considered significantly different [48]. The Nemenyi test calculates a critical difference (CD) threshold:

$$CD = q_\alpha \times \sqrt{\frac{k(k+1)}{6N}} \quad (6)$$

Here, q_α is the critical value obtained from the Studentized range distribution, which depends on the number of classifiers and the chosen significance level (typically $\alpha = 0.05$). The terms k and N are as previously defined. If the difference in average ranks between any two classifiers exceeds the critical difference, their performances are considered statistically different.

3. RESULT

This section presents the evaluation results of six machine learning classifiers, namely Random Forest (RF), K-Nearest Neighbors (KNN), Naive Bayes (NB), Multilayer Perceptron (MLP), Support Vector Machine (SVM), and Hidden Markov Model (HMM), across four SIBI gesture categories, which are Affix, Alphabet, Number, and Word. Each model was trained and validated using 5-fold stratified cross-validation, and their performance was assessed using accuracy, precision, recall, and F1-score. To ensure the reliability of the findings and validate whether the observed differences in classifier performance were statistically significant, the Friedman test was applied, followed by the Nemenyi post-hoc test and Critical Difference (CD) diagrams.

3.1. Overall Classifier Performance

The performance of six traditional classifiers (Random Forest (RF), K-Nearest Neighbors (KNN), Naïve Bayes (NB), Multilayer Perceptron (MLP), Support Vector Machine (SVM), and Hidden Markov Model (HMM)) across the four SIBI gesture categories is summarized in Table 1, Figure 2 and Figure 3. The reported metrics include training accuracy, validation accuracy, precision, recall, and F1-score, each expressed as mean ± standard deviation over five folds. In Figures 2 and 3, the vertical bars on top of each column represent the standard deviation (SD), indicating the variability of the metric across folds. Shorter SD bars suggest the model's performance is more consistent across different data splits, while longer bars indicate greater variability. Overall, non-linear models such as MLP and RF demonstrated superior generalization across all gesture types, while NB consistently exhibited the weakest performance due to its independence assumption, which does not align with the highly correlated, 256-dimensional BiLSTM features used as input.

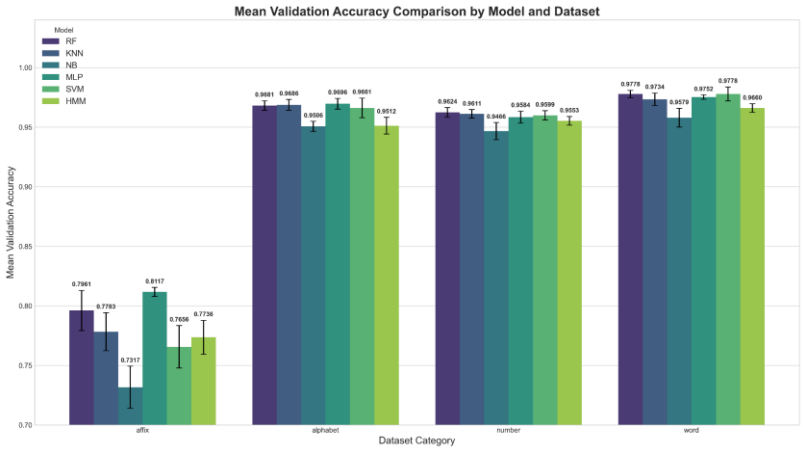


Figure 2. Mean validation accuracy (with standard deviation) of all classifiers across four SIBI gesture categories

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No. 2, April 2026, Page. 823-842
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5210

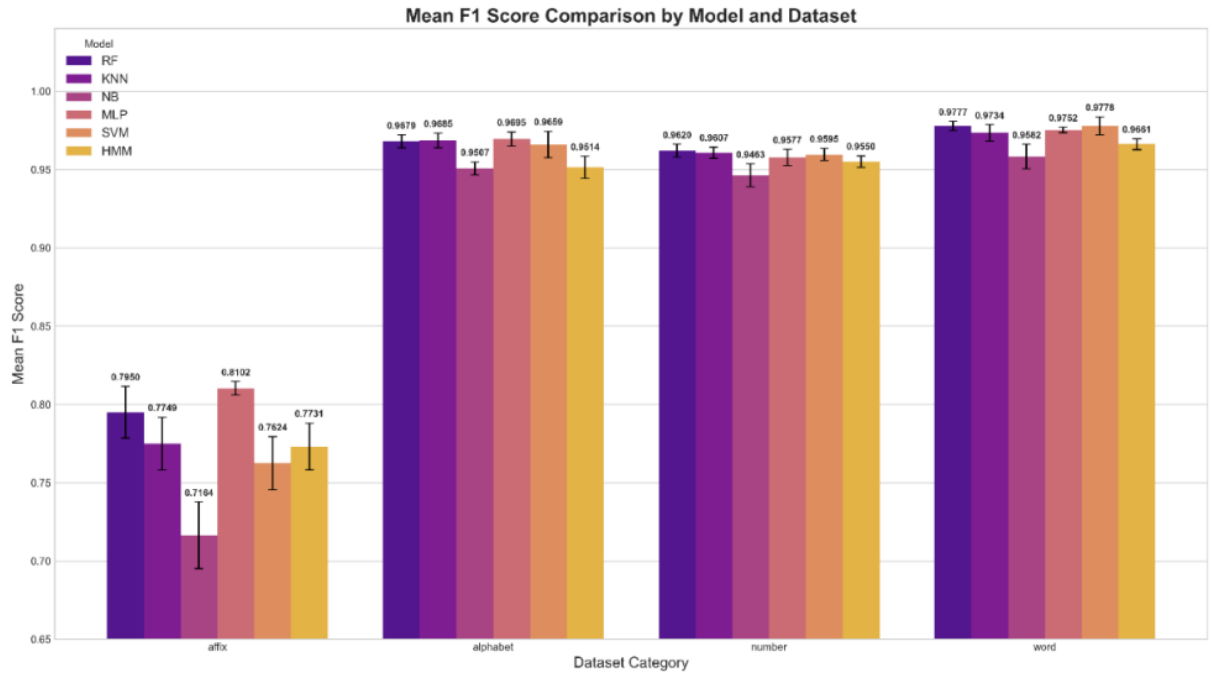


Figure 3. Mean F1 Score (with standard deviation) of all classifiers across four SIBI gesture categories

For Affix gestures, which represent the most complex category due to subtle visual variations, validation accuracy ranged from 0.7317 (NB) to 0.8117 (MLP). Among all models, MLP achieved the highest validation accuracy (0.8117 ± 0.0038) and F1-score (0.8102 ± 0.0044), with a very small standard deviation, indicating stable performance despite the category's subtle motion variations. RF achieved a slightly lower F1-score of 0.7950 ± 0.0165 but with a higher standard deviation (0.0169), while KNN, HMM, and SVM recorded mid-range values between 0.762 and 0.775. NB not only recorded the lowest mean values but also exhibited larger standard deviation, reflecting both lower accuracy and less stability.

For Alphabet gestures, all classifiers achieved robust performance, with validation accuracy exceeding 0.95 for all except NB (0.9506 ± 0.0042). MLP again recorded the highest results, with a validation accuracy of 0.9696 ± 0.0045 and an F1-score of 0.9695 ± 0.0045. Its performance was only slightly higher than that of KNN and RF, which both achieved approximately 0.968. The small standard deviations (≤ 0.005) for the top-performing models indicate that their high performance was consistent across all cross-validation folds, with minimal variability.

For Number gestures, validation accuracy remained high across all classifiers, ranging from 0.9466 (NB) to 0.9624 (RF). RF achieved the highest validation accuracy of 0.9624 ± 0.0040 and an F1-score of 0.9620 ± 0.0041, closely followed by KNN at 0.9607 ± 0.0036 and SVM at 0.9595 ± 0.0039. MLP and HMM followed with slightly lower scores, while NB again ranked last. The very low standard deviations for RF, KNN, and SVM (≤ 0.004) show that these models not only performed well but also maintained stability across different folds.

For Word gestures, all models demonstrated strong results, though the differences between the top and bottom classifiers were more noticeable. RF achieved perfect training accuracy of 1.0000 ± 0.0000 and the highest validation accuracy of 0.9778 ± 0.0031 and F1-score of 0.9777 ± 0.0030. SVM and MLP closely followed, with F1-scores of 0.9778 ± 0.0058 and 0.9752 ± 0.0018, respectively. Almost all model achieved the result with very small standard deviations. However, NB achieved a validation accuracy of 0.9579 ± 0.0079 and an F1-score of 0.9582 ± 0.0078, which was notably lower than the other models. NB also have a larger standard deviation (0.0079) compared to the top models.

Vol. 7, No. 2, April 2026, Page. 823-842
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5210

Table 1. Performance Metrics (Mean ± Standard Deviation) of Six Classifiers on BiLSTM Features for SIBI Recognition

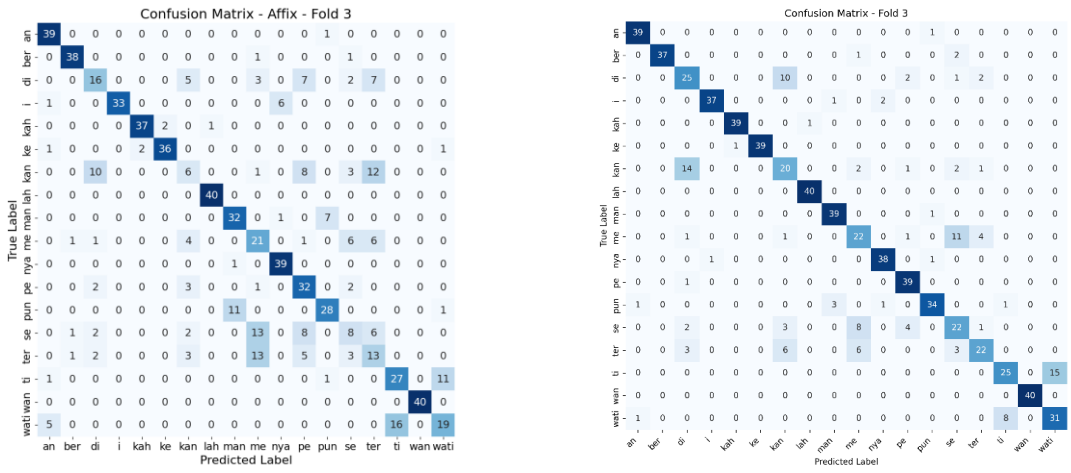
Category	ML Model	Training Accuracy	Validation Accuracy	Precision	Recall	F1 Score
Affix	RF	0.9978 ± 0.0007	0.7961 ± 0.0169	0.7965 ± 0.0156	0.7961 ± 0.0169	0.7950 ± 0.0165
	KNN	0.8538 ± 0.0035	0.7783 ± 0.0159	0.7778 ± 0.0178	0.7783 ± 0.0159	0.7749 ± 0.0169
	NB	0.7603 ± 0.0022	0.7317 ± 0.0176	0.7401 ± 0.0201	0.7317 ± 0.0176	0.7164 ± 0.0213
	MLP	0.9308 ± 0.0187	0.8117 ± 0.0038	0.8198 ± 0.0085	0.8117 ± 0.0038	0.8102 ± 0.0044
	SVM	0.7803 ± 0.0038	0.7656 ± 0.0178	0.7633 ± 0.0174	0.7656 ± 0.0178	0.7624 ± 0.0170
	HMM	0.8053 ± 0.0035	0.7736 ± 0.0142	0.7764 ± 0.0149	0.7736 ± 0.0142	0.7731 ± 0.0150
Alphabet	RF	0.9946 ± 0.0006	0.9681 ± 0.0040	0.9686 ± 0.0042	0.9680 ± 0.0041	0.9679 ± 0.0041
	KNN	0.9755 ± 0.0014	0.9686 ± 0.0046	0.9692 ± 0.0043	0.9687 ± 0.0046	0.9685 ± 0.0047
	NB	0.9622 ± 0.0009	0.9506 ± 0.0042	0.9525 ± 0.0030	0.9507 ± 0.0043	0.9507 ± 0.0042
	MLP	0.9921 ± 0.0013	0.9696 ± 0.0045	0.9705 ± 0.0041	0.9696 ± 0.0045	0.9695 ± 0.0045
	SVM	0.9706 ± 0.0017	0.9661 ± 0.0083	0.9671 ± 0.0081	0.9662 ± 0.0083	0.9659 ± 0.0084
	HMM	0.9720 ± 0.0020	0.9512 ± 0.0070	0.9538 ± 0.0061	0.9511 ± 0.0070	0.9514 ± 0.0071
Number	RF	0.9975 ± 0.0005	0.9624 ± 0.0040	0.9631 ± 0.0039	0.9621 ± 0.0041	0.9620 ± 0.0041
	KNN	0.9735 ± 0.0009	0.9611 ± 0.0036	0.9619 ± 0.0035	0.9606 ± 0.0036	0.9607 ± 0.0036
	NB	0.9590 ± 0.0016	0.9466 ± 0.0072	0.9485 ± 0.0067	0.9464 ± 0.0074	0.9463 ± 0.0073
	MLP	0.9794 ± 0.0019	0.9584 ± 0.0050	0.9598 ± 0.0049	0.9580 ± 0.0051	0.9577 ± 0.0052
	SVM	0.9635 ± 0.0010	0.9599 ± 0.0039	0.9612 ± 0.0038	0.9594 ± 0.0040	0.9595 ± 0.0039
	HMM	0.9686 ± 0.0008	0.9553 ± 0.0037	0.9564 ± 0.0036	0.9550 ± 0.0037	0.9550 ± 0.0037
Word	RF	1.0000 ± 0.0000	0.9778 ± 0.0031	0.9784 ± 0.0028	0.9778 ± 0.0031	0.9777 ± 0.0030
	KNN	0.9818 ± 0.0015	0.9734 ± 0.0052	0.9742 ± 0.0049	0.9734 ± 0.0052	0.9734 ± 0.0052
	NB	0.9676 ± 0.0020	0.9579 ± 0.0079	0.9606 ± 0.0074	0.9579 ± 0.0079	0.9582 ± 0.0078
	MLP	0.9976 ± 0.0005	0.9752 ± 0.0018	0.9758 ± 0.0016	0.9752 ± 0.0018	0.9752 ± 0.0018
	SVM	0.9825 ± 0.0006	0.9778 ± 0.0059	0.9785 ± 0.0058	0.9778 ± 0.0059	0.9778 ± 0.0058
	HMM	0.9844 ± 0.0013	0.9660 ± 0.0036	0.9674 ± 0.0035	0.9660 ± 0.0036	0.9661 ± 0.0035

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No. 2, April 2026, Page. 823-842
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5210

3.2. Confusion Matrix Analysis

The Affix category remains the most challenging subset of SIBI because many of its gestures differ only in minute details, often the precise placement of a thumb, a slight bend of a finger, or a barely perceptible hand rotation, while sharing nearly identical orientations and trajectories. Pairs such as “se” and “me” illustrate this difficulty, both gestures involve two hands, yet “me” hides the thumb inside the fist whereas “se” keeps it outside. A similarly subtle distinction separates “ter” from “me,” where the thumb is tucked beneath the index finger in “ter,” but the overall motion is otherwise alike. Other confusable pair include “ti” and “wati”, which differ only by small rotational adjustments or modest changes in finger extension.

In the first study, a BiLSTM network trained end-to-end struggled with precisely these fine-grained distinctions, and the resulting confusion matrix showed sizeable off-diagonal counts concentrated in the ambiguous pairs. By replacing the BiLSTM’s final soft-max layer with a Multilayer Perceptron that operates on the same 256-dimensional embeddings, the present work achieves a clearer separation among these difficult classes. Misclassifications between “se,” “me,” and “ter,” which previously appeared in double-digit counts, now drop to smaller number figures, and the spill-over between “ti,” “wan,” and “wati” is roughly halved.



(a) (b)

Figure 4. Comparison of aggregate best fold confusion matrices for Affix gestures between (a) BiLSTM and (b) BiLSTM + MLP

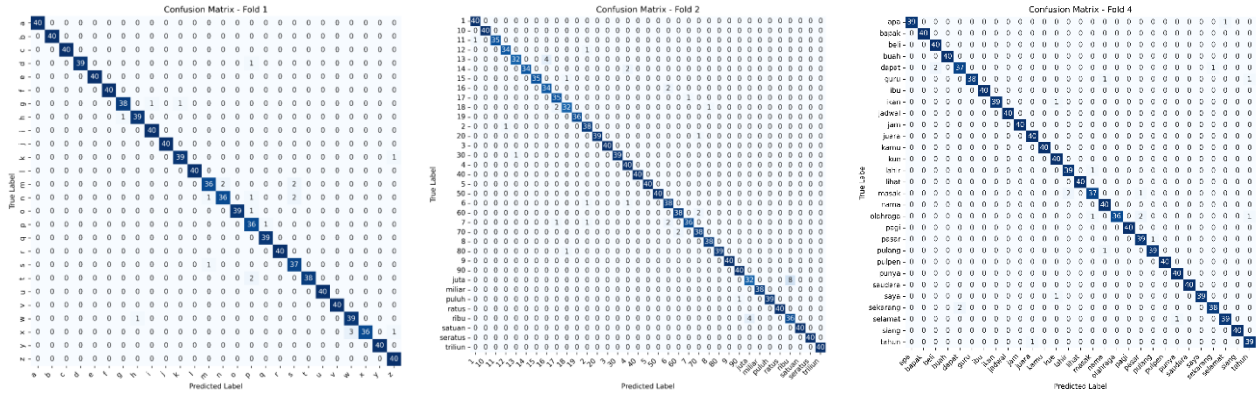
The confusion matrix for the Alphabet category, utilizing the MLP classifier, shows near-perfect performance, as evident from the predominantly dark diagonal indicating high correct predictions for each letter. Almost all letters, including "a," "b," "c," "d," "e," "f," and so forth, were accurately predicted with minimal or no confusion. However, a few minor misclassifications occurred, for example between the letters "m" and "n," which is likely due to their similar hand shapes differing only slightly in finger orientation. However in this case, the error rate for both gesture is decreasing. The minimal confusion underscores the distinct and isolated nature of alphabet gestures, which involve static hand shapes without significant temporal variation.

For the Number category, the confusion matrix from the Random Forest (RF) classifier similarly demonstrates strong performance, with clear diagonal dominance indicating high recognition accuracy for numeric gestures. The model was particularly successful in identifying distinct numeric labels such as "1," "10," "20," "30," "40," and higher numerical scales like "juta" (million) and "ribu" (thousand), with minimal confusion. Nonetheless, certain numeric classes exhibited some misclassification, particularly among numbers like "13," "14," and "15," reflecting the visual similarity inherent to gestures representing

Vol. 7, No. 2, April 2026, Page. 823-842
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5210

sequential numeric values. This pattern of confusion is consistent with numeric gestures' sequential nature, where gestures differ subtly in finger orientation or movement direction.

The confusion matrix for the Word category (using the RF classifier) again displays very high performance with predominant diagonal values, indicating correct classifications across nearly all word classes such as "apa," "bapak," "beli," "buah," and "guru." The accuracy is consistently high, with only minor errors scattered across the matrix and these errors remain minimal.



(a) (b) (c)

Figure 5. Aggregate confusion matrix of the best-performing classifier (MLP for Alphabet, RF for Number, RF for Word) averaged across five folds.

3.3. Statistical Tests

To strengthen the validity of our classifier evaluation and go beyond raw performance metrics (accuracy and F1-score), post-hoc statistical analyses using the Nemenyi test were applied. This test compares all classifiers pairwise based on their average ranks, obtained across five stratified cross-validation folds. The results are visualized through two complementary tools, which are Critical Difference (CD) diagrams and Nemenyi pairwise comparison heatmaps.

The resulting Chi-Squared values for all categories, which are Affix (19.74), Alphabet (17.87), Number (11.16), and Word (20.20), all exceeded the critical threshold of 11.0705 for 6 classifiers (df = 5, $\alpha = 0.05$), with corresponding p-values below 0.05. These results confirm that the observed performance differences are statistically significant and not due to random variation.

Table 2. Chi-Squared and p-values from the Friedman Test Across Gesture Categories

Category	Chi-Squared	p-value
Affix	19.74	0.0014
Alphabet	17.87	0.0031
Number	11.16	0.0482
Word	20.20	0.0011

In Nemenyi post-hoc, the value between 0 and 1 is the p-value. The lower it is, the stronger the evidence that one model is significantly different from the other. The Nemenyi post-hoc heatmap for the Affix dataset revealed significant pairwise differences in performance. Notably, MLP significantly outperformed NB ($p = 0.0003$) and SVM ($p = 0.0467$). No other comparisons were statistically significant, indicating that the remaining models such as RF, KNN, and HMM, did not differ significantly from one another or from MLP in terms of average rank. This suggests that while MLP clearly surpassed NB and showed a slight edge over SVM, the differences among the mid-performing models were not statistically strong.

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No. 2, April 2026, Page. 823-842
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5210

In the Alphabet dataset, the Nemenyi heatmap shows that MLP significantly outperformed NB ($p = 0.0284$). While KNN and RF had strong average ranks, their differences with MLP were not statistically significant ($p \approx 0.9999$). Similarly, no significant differences were found among KNN, RF, and SVM. This implies that although MLP, KNN, and RF were the top performers, only the performance gap between MLP and NB reached statistical significance.

For the Number dataset, the Friedman test was barely significant ($p = 0.0482$), and this is reflected in the Nemenyi heatmap, where none of the pairwise comparisons reached statistical significance at $p < 0.05$, except for a borderline case between RF and NB ($p = 0.0592$). This suggests that although RF had the best average rank, the differences were not statistically strong enough to confidently distinguish it from the other models. The performance gaps in this category should be interpreted with caution due to their marginal significance.

In the Word dataset, several statistically significant pairwise differences were observed. RF significantly outperformed NB ($p = 0.0028$) and HMM ($p = 0.0467$). Additionally, SVM also outperformed NB ($p = 0.0284$). No other comparisons were statistically significant. These results confirm that RF and SVM were consistently strong performers, while NB and HMM lagged significantly behind. MLP, despite ranking third, did not show significant differences from its neighboring models.

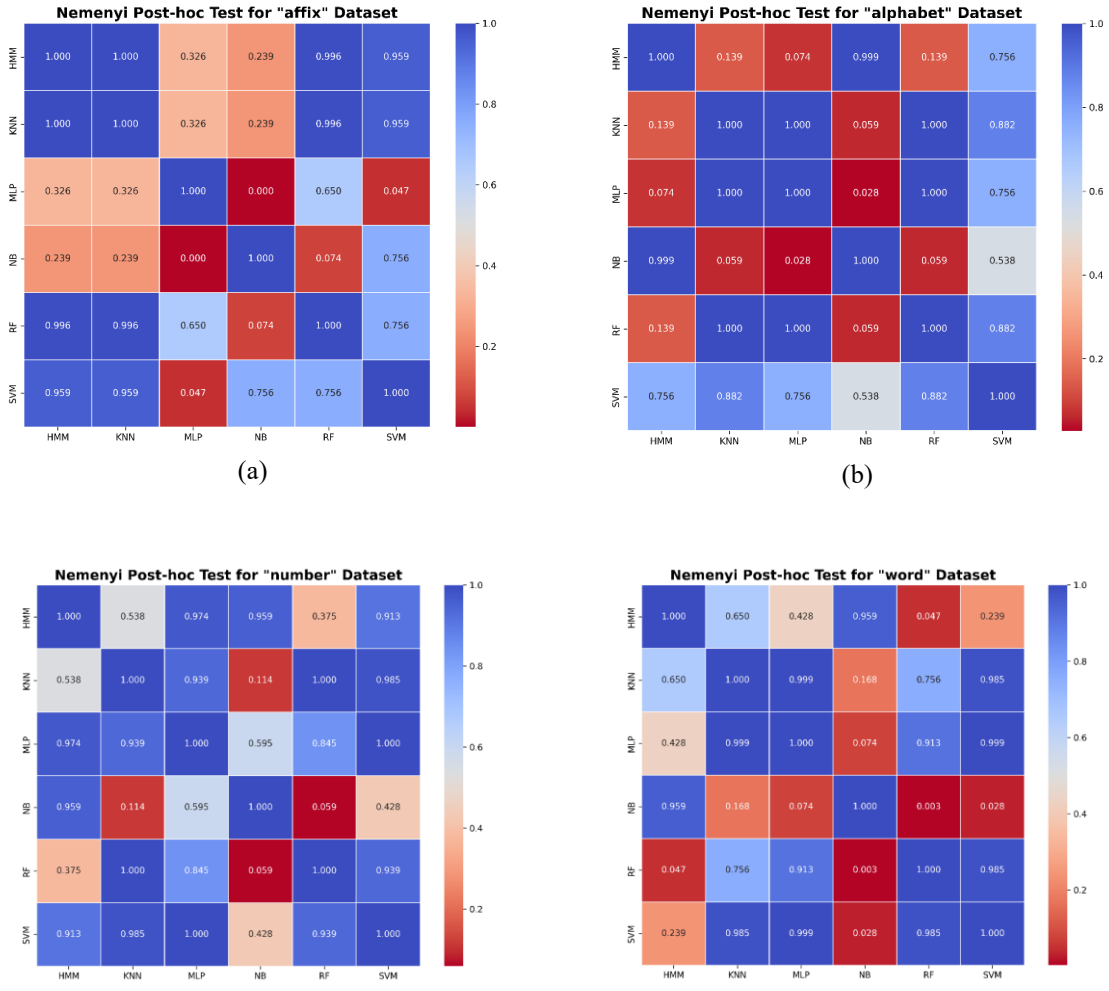


Figure 6. Nemenyi post-hoc heatmaps comparing pairwise p-values between classifiers across all gesture categories: (a) Affix, (b) Alphabet, (c) Number, and (d) Word.

Vol. 7, No. 2, April 2026, Page. 823-842
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5210

Despite the significant findings from the Nemenyi heatmaps, the Critical Difference (CD) diagrams in Figure 7 show that all classifiers fall within a single connected group across all four gesture categories. This visualization represents model groupings based on the Nemenyi test's critical difference (CD) threshold, which defines the minimum difference in average ranks required for the performance of two models to be considered statistically distinct. In each CD diagram, models are placed on a horizontal axis according to their average rank (lower is better), and classifiers that are not significantly different from each other are connected with a horizontal line.

The presence of a single group in each diagram indicates that, although certain model pairs, such as MLP vs NB in Affix or RF vs NB in Word, showed statistically significant differences in pairwise comparisons, the overall differences across all models were not large enough to exceed the CD threshold. This outcome can be attributed to the conservative nature of the Nemenyi test, combined with the limited number of cross-validation folds (n = 5), which reduces the test's sensitivity. Additionally, it reflects that while some models consistently performed better (MLP, RF), the magnitude of performance gaps among the mid-ranked classifiers (KNN, HMM, SVM) was relatively small and statistically indistinct.

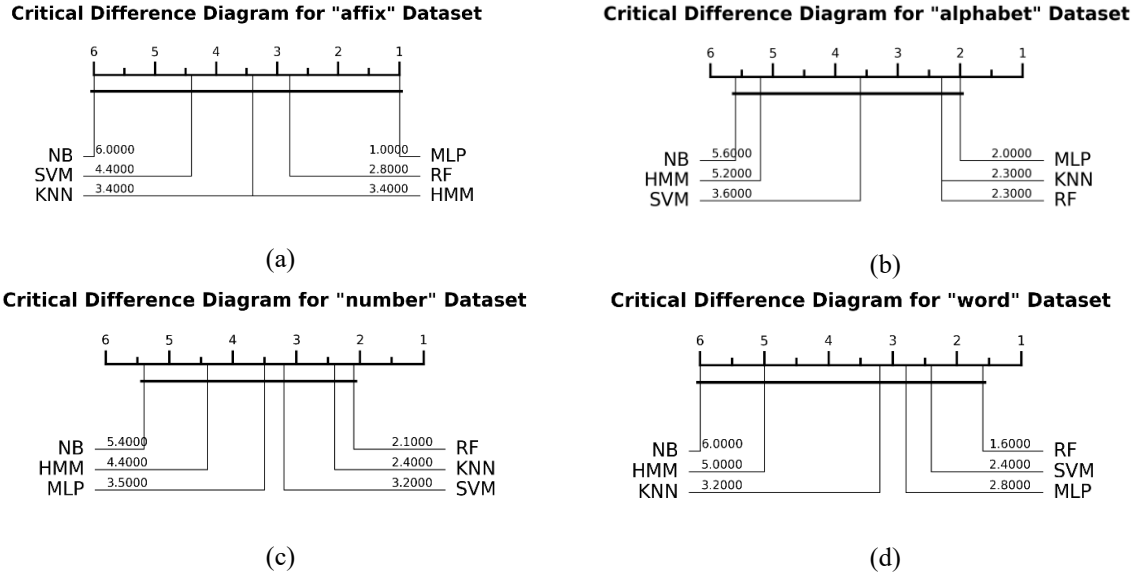


Figure 7. Critical Difference diagrams illustrating average ranks and statistical groupings of classifiers across each gesture category: (a) Affix, (b) Alphabet, (c) Number, and (d) Word.

4. DISCUSSIONS

This study presents an extended analysis of hand gesture recognition for the Indonesian Sign Language System (SIBI), building upon our previously published LSTM-based approach. Unlike the earlier work, which focused solely on deep learning through a BiLSTM architecture, the current experiment evaluates a diverse set of traditional and hybrid classifiers, which are Random Forest (RF), K-Nearest Neighbors (KNN), Naive Bayes (NB), Multilayer Perceptron (MLP), Support Vector Machine (SVM), and Hidden Markov Model (HMM), using features extracted from MediaPipe keypoints. This comparative analysis is designed to assess the viability of these models as lighter and more interpretable alternatives to deep neural networks.

The overall performance trend remains consistent with prior findings. As shown in Table 1, the Affix category continues to be the most challenging across all models, with the best validation accuracy reaching only 81.17% (MLP), while Alphabet, Number, and Word categories achieve considerably higher validation accuracies, often exceeding 96%. This aligns with earlier observations, where BiLSTM achieved mean

Vol. 7, No. 2, April 2026, Page. 823-842
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5210

accuracies of 68.17%, 93.94%, 91.48%, and 92.41% for Affix, Alphabet, Number, and Word, respectively [1]. These results reaffirm that although there is a degree of improvement in recognizing affix gestures, they remain visually ambiguous and difficult to distinguish due to their subtle variations in thumb placement or finger curl ("me" vs. "se" or "ter"), making them still challenging even for human observers.

A real-time Indonesian sign language study using Mediapipe Holistic landmarks and an LSTM evaluates two settings, which are a smaller dataset without k-fold training that reaches about 85% accuracy, and a larger dataset with 5-fold training at 30 epochs per fold that averages about 98% accuracy. The dataset covers six everyday words with 80 videos per word and includes live trials on three subjects with nine attempts each, reporting about 92% detection and noting occasional confusions for visually similar motions [17]. Relative to that setup, our experiment focus on four SIBI categories (Alphabet, Number, Word, Affix), uses frozen LSTM embeddings as features for six lightweight classifiers, applies 5-fold stratified cross-validation across all classifiers, and reports both accuracy and macro-F1 along with significance testing. A separate SIBI study builds an end-to-end video model with an Inflated 3D CNN initialized from large-scale priors; with 200 videos covering 10 words and 2 signers, the best configuration attains 97.5% testing accuracy by freezing earlier inception modules [49]. In contrast, our pipeline operates on keypoint-derived embeddings, compares six classical classifiers under identical folds, and emphasizes statistical analysis of average ranks. Another experiment using Indian sign language paper centers on an SVM-based recognition pipeline [50]; in our case, SVM is evaluated alongside KNN, RF, MLP, NB, and HMM on the same frozen-embedding features and shared folds, enabling a within-dataset comparison across algorithms.

To validate performance differences beyond raw metrics, we conducted Friedman tests followed by Nemenyi post-hoc analysis across all four gesture categories. The Friedman test yielded statistically significant results in all cases ($p < 0.05$), confirming meaningful performance differences across classifiers. However, despite some significant pairwise results in the Nemenyi heatmaps, such as MLP significantly outperforming NB in the Affix and Alphabet datasets, none of the classifier pairs exceeded the critical difference threshold in the CD diagrams. Consequently, all models appeared as a single group in each diagram. This suggests that while individual model differences exist, their performance variations are not large enough to be considered statistically distinct when correcting for multiple comparisons.

Interestingly, MLP and RF frequently achieved top validation scores across categories, indicating their ability to model complex feature interactions effectively. NB consistently underperformed, as expected, due to its strong independence assumption which is poorly aligned with the correlated nature of sequential keypoint features. KNN and HMM, while not always top performers, maintained stable results, reflecting their strength in sequence modeling.

Compared to the BiLSTM approach in our previous work, traditional models, especially MLP and RF, offer surprisingly competitive performance, particularly in Alphabet, Number, and Word categories. However, their accuracy in the Affix category remains limited, reinforcing that subtle spatial-temporal patterns in affix gestures may require deeper temporal context and representation learning, which BiLSTM inherently captures. These findings provide a practical guideline for computer science and informatics, which are lightweight models such as MLP and RF can be preferred for visually distinct gesture classes due to their efficiency and interpretability, while categories with fine-grained variations benefit from deeper temporal architectures. Moreover, the consistent difficulty of Affix gestures offers a benchmark for future research on richer spatiotemporal representations, targeted data augmentation, and improved labeling consistency.

5. CONCLUSION

This study expands upon previous work on Indonesian Sign Language System (SIBI) gesture recognition by evaluating a range of traditional and hybrid classifiers, namely RF, KNN, NB, MLP, SVM,

and HMM, using keypoint features extracted from MediaPipe. Performance was assessed across four SIBI categories, which are Affix, Alphabet, Number, and Word, with evaluation using 5-fold cross-validation, followed by statistical validation through Friedman and Nemenyi post-hoc tests.

The results show that MLP and RF consistently achieved strong performance in Alphabet, Number, and Word categories, with validation accuracies exceeding 96%. In the Affix category, MLP reached 81.17%, marking a substantial improvement over the 68.17% achieved by our earlier BiLSTM model, although affix gestures remain more challenging due to minimal visual differences in hand shapes and movements.

Statistical testing confirmed that the observed performance differences between classifiers were significant ($p < 0.05$). However, the Nemenyi post-hoc test and Critical Difference diagrams revealed that most classifiers were statistically tied, indicating that the performance gaps, while measurable, were not large enough to be considered distinct when adjusting for multiple comparisons.

In comparison to our previous BiLSTM-based approach, traditional models, particularly MLP and RF demonstrated competitive performance in several categories. However, BiLSTM remains more effective in handling subtle temporal dependencies, especially for visually similar affix gestures. Future research should explore incorporating additional input modalities beyond keypoints, such as hand segmentation masks, hand shape contours, or RGB hand crops, to provide richer spatial information. These visual cues, combined with non-manual features like facial expressions, may help the model better understand subtle gesture variations, especially in visually similar affix signs. Additionally, exploring joint modeling of affix and word categories could allow the system to learn shared linguistic structures and improve disambiguation in more complex gesture combinations.

CONFLICT OF INTEREST

The authors declares that there is no conflict of interest between the authors or with research object in this paper.

REFERENCES

[1] P. Ho and H. Santoso, "LSTM-Based Hand Gesture Recognition for Indonesian Sign Language System (SIBI) on Affix, Alphabet, Number, and Word," *Journal of Applied Informatics and Computing (JAIC)*, vol. 9, no. 3, pp. 928–937, Jun. 2025, doi: 10.30871/jaic.v9i3.9607.

[2] K. Gajurel, C. Zhong, and G. Wang, "A Fine-Grained Visual Attention Approach for Fingerspelling Recognition in the Wild," in *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Melbourne: IEEE, May 2021, pp. 3266–3271. doi: 10.1109/SMC52423.2021.9658982.

[3] 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.

[4] 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*, Badung: Institute of Electrical and Electronics Engineers Inc., Nov. 2023, pp. 23–28. doi: 10.1109/ICSGTEIS60500.2023.10424020.

[5] F. Bigand, E. Prigent, B. Berret, and A. Braffort, "Machine Learning of Motion Statistics Reveals the Kinematic Signature of the Identity of a Person in Sign Language," Frontiers in Bioengineering and Biotechnology, vol. 9, pp. 1–10, Jul. 2021, doi: 10.3389/fbioe.2021.710132.

[6] A. Safonova, G. Ghazaryan, S. Stiller, M. Main-Knorn, C. Nendel, and M. Ryo, "Ten deep learning techniques to address small data problems with remote sensing," *International Journal of Applied*

Vol. 7, No. 2, April 2026, Page. 823-842
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5210

Earth Observation and Geoinformation, vol. 125, p. 103569, Dec. 2023, doi: 10.1016/J.JAG.2023.103569.

[7] I. H. Rather, S. Kumar, and A. H. Gandomi, "Breaking the data barrier: a review of deep learning techniques for democratizing AI with small datasets," Artificial Intelligence Review, vol. 57, no. 9, p. 226, Aug. 2024, doi: 10.1007/S10462-024-10859-3.

[8] W. Chen, K. Yang, Z. Yu, Y. Shi, and C. L. P. Chen, "A Survey on Imbalanced Learning: Latest Research, Applications and Future Directions," Artificial Intelligence Review, vol. 57, no. 6, p. 137, Jun. 2024, doi: 10.1007/S10462-024-10759-6/FIGURES/11.

[9] O. Khattach, O. Moussaoui, and M. Hassine, "Enhancing Machine Failure Prediction with a Hybrid Model Approach," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 13, no. 3, pp. 2946–2955, Sep. 2024, doi: 10.11591/ijai.v13.i3.pp2946-2955.

[10] M. A. Metu, N. Akhter, S. Nasrin, T. Anzum, A. Khatun, and R. Mazumder, "Hybrid SVM-Bidirectional Long Short-Term Memory Model for Fine-Grained Software Requirement Classification," *Journal of Advances in Information Technology*, vol. 15, no. 8, pp. 914–922, Aug. 2024, doi: 10.12720/jait.15.8.914-922.

[11] S. Kamble, "SLRNet: A Real-Time LSTM-Based Sign Language Recognition System," Jun. 2025. doi: 10.48550/arXiv.2506.11154.

[12] S. Djaballah, L. Saidi, K. Meftah, A. Hechifa, M. Bajaj, and I. Zaitsev, "A Hybrid LSTM Random Forest Model with Grey Wolf Optimization for Enhanced Detection of Multiple Bearing Faults," Scientific Reports, vol. 14, p. 23997, Dec. 2024, doi: 10.1038/s41598-024-75174-x.

[13] M. Haidarh, C. Mu, Y. Liu, and X. He, "Exploring Traditional, Deep Learning and Hybrid Methods for Hyperspectral Image Classification: A Review," *Journal of Information and Intelligence*, Apr. 2025, doi: 10.1016/J.JIIXD.2025.04.002.

[14] C. Haibin and H. Yongliang, "A Hybrid LSTM and Decision Tree Model: A Novel Machine Learning Architecture for Complex Data Classification," in *2023 IEEE International Conference on Sensors, Electronics and Computer Engineering*, Jinzhou: IEEE, Aug. 2023, pp. 1441–1446. doi: 10.1109/ICSECE58870.2023.10263317.

[15] L. Li, Y. Wu, Y. Ou, Q. Li, Y. Zhou, and D. Chen, "Research on Machine Learning Algorithms and Feature Extraction for Time Series," in *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, Montreal: IEEE, Jul. 2017, pp. 1–5. doi: 10.1109/PIMRC.2017.8292668.

[16] 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," Advanced Science Letters, vol. 24, no. 2, pp. 999–1004, Mar. 2018, doi: 10.1166/ASL.2018.10675.

[17] C. A. Sari, E. H. Rachmawanto, Z. Saifullah, C. Jatmoko, D. Sinaga, and S. Program, "Real-Time Detection of Indonesian Sign Language (ISL) Gestures Based on Long Short-Term Memory," *Journal of Soft Computing Exploration*, vol. 5, no. 3, pp. 251–262, Sep. 2024, doi: 10.52465/JOSCEX.V5I3.452.

[18] H. Joshi, V. Golhar, J. Gundawar, A. Gangurde, A. Yenkikar, and N. P. Sable, "Real-Time Sign Language Recognition and Sentence Generation," in *Proceedings of the International Conference on Innovative Computing & Communication (ICICC 2024)*, Elsevier BV, Oct. 2024. doi: 10.2139/SSRN.4992818.

[19] J. P. Sahoo, S. Ari, and S. K. Patra, "Hand Gesture Recognition using PCA based Deep CNN Reduced Features and SVM Classifier," in *Proceedings - 2019 IEEE International Symposium on Smart Electronic Systems, iSES 2019*, Rourkela: IEEE, Dec. 2019, pp. 221–224. doi: 10.1109/ISES47678.2019.00056.

[20] N. Zerrouki, A. Houacine, F. Harrou, R. Bouarroudj, M. Y. Cherifi, and Y. Sun, "Exploiting Deep Learning-Based LSTM Classification for Improving Hand Gesture Recognition to Enhance

Vol. 7, No. 2, April 2026, Page. 823-842
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5210

Visitors' Museum Experiences," in *2022 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies, 3ICT 2022*, Sakheer: Institute of Electrical and Electronics Engineers Inc., Nov. 2022, pp. 451–456. doi: 10.1109/3ICT56508.2022.9990722.

[21] R. E. Nogales and M. E. Benalcázar, "Hand Gesture Recognition Using Automatic Feature Extraction and Deep Learning Algorithms with Memory," *Big Data and Cognitive Computing*, vol. 7, no. 2, p. 102, May 2023, doi: 10.3390/BDCC7020102.

[22] S. Wu, "A Compact LSTM-SVM Fusion Model for Long-Duration Cardiovascular Diseases Detection," Michigan, Nov. 2023. doi: 10.48550/arXiv.2312.09442.

[23] Z. F. Jailani and D. Nurmadewi, "Hybrid Machine Learning Predicts Flooding Using LSTM And Random Forests On Geodata," *INTECOMS: Journal of Information Technology and Computer Science*, vol. 8, no. 1, pp. 35–41, Jan. 2025, doi: 10.31539/INTECOMS.V8I1.13991.

[24] I. Simatupang, D. S. Pamungkas, and S. K. Risandriya, "Naïve Bayes Classifier for Hand Gestures Recognition," in *Proceedings of the 3rd International Conference on Applied Engineering - ICAE*, Batam: Scitepress, Jul. 2021, pp. 110–114. doi: 10.5220/0010352601100114.

[25] H. Ma', A. Yudo Husodo, and B. Irmawati, "Performance Comparison of Naive Bayes and Bidirectional LSTM Algorithms In BSI Mobile Review Sentiment Analysis," *Jurnal Teknik Informatika (JUTIF)*, vol. 6, no. 1, pp. 159–172, Feb. 2025, doi: 10.52436/1.jutif.2025.6.1.4178.

[26] S. Widodo, H. Brawijaya, and S. Samudi, "Stratified K-fold Cross Validation Optimization on Machine Learning for Prediction," *Sinkron : Jurnal dan Penelitian Teknik Informatika*, vol. 6, no. 4, pp. 2407–2414, Oct. 2022, doi: 10.33395/SINKRON.V7I4.11792.

[27] 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," Aug. 13, 2024, *Mendeley Data*. doi: 10.17632/44PBRBSNKH.3.

[28] A. R. Isnain, J. Supriyanto, and M. P. Kharisma, "Implementation of K-Nearest Neighbor (K-NN) Algorithm For Public Sentiment Analysis of Online Learning," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 15, no. 2, pp. 121–130, Apr. 2021, doi: 10.22146/IJCCS.65176.

[29] S. Shivani and S. B. Gupta, "A Comprehensive Analysis of Recognition of Hand Gestures using Machine Learning," *Makara Journal of Technology*, vol. 29, no. 1, p. 5, Apr. 2025, doi: 10.7454/mst.v29i1.1679.

[30] A. Desiani *et al.*, "Penerapan Metode Support Vector Machine Dalam Klasifikasi Bunga Iris," *IJAI (Indonesian Journal of Applied Informatics)*, vol. 7, no. 1, pp. 12–18, Apr. 2023, doi: 10.20961/IJAI.V7I1.61486.

[31] S. Sharma, S. Modi, P. S. Rana, and J. Bhattacharya, "Hand Gesture Recognition Using Gaussian Threshold and Different SVM Kernels," in *International Conference on Advances in Computing and Data Sciences*, M. Singh, P. K. Gupta, V. Tyagi, J. Flusser, and T. Ören, Eds., Dehradun: Springer, Singapore, Oct. 2018, pp. 138–147. doi: 10.1007/978-981-13-1813-9_14.

[32] K. Nguyen-Trong and T. T. T. Nguyen, "Optimization of Multi-Layer Perceptron Deep Neural Networks using Genetic Algorithms for Hand Gesture Recognition," *Journal of Computer Science*, vol. 18, no. 2, pp. 57–66, Feb. 2022, doi: 10.3844/jcssp.2022.57.66.

[33] D. Pardede, B. H. Hayadi, and Iskandar, "Kajian Literatur Multi Layer Perceptron: Seberapa Baik Performa Algoritma Ini," *Journal of ICT Aplication and System*, vol. 1, no. 1, pp. 23–35, Jun. 2022, doi: 10.56313/jictas.v1i1.127.

[34] N. Neverova, C. Wolf, G. W. Taylor, and F. Nebout, "Multi-scale Deep Learning for Gesture Detection and Localization," in *Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, L. Agapito, M. M. Bronstein, and C. Rother, Eds., Zurich: Springer, Cham, Jan. 2015, pp. 474–490. doi: 10.1007/978-3-319-16178-5_33.

Vol. 7, No. 2, April 2026, Page. 823-842
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5210

[35] N. S. O'Connell, B. C. Jaeger, G. S. Bullock, and J. L. Speiser, "A Comparison of Random Forest Variable Selection Methods for Regression Modeling of Continuous Outcomes," Briefings in Bioinformatics, vol. 26, no. 2, pp. 1–15, Mar. 2025, doi: 10.1093/BIB/BBAF096.

[36] A. Chauhan and E. Shyni, "Affordable Real-Time Hand Gesture Detection Using Random Forest," International Journal of Innovative Science and Research Technology, vol. 10, no. 2, pp. 183–190, Feb. 2025, doi: 10.5281/zenodo.14898697.

[37] S. Pavan and Y. Gowtham, "Hand Gesture Recognition System Using Random Forest," *International Journal of Advance Research in Science and Engineering*, vol. 12, no. 4, pp. 125–129, Apr. 2023, Accessed: Jul. 29, 2025. [Online]. Available: https://www.ijarse.com/images/fullpdf/1682502367_737.pdf

[38] O. Peretz, M. Koren, and O. Koren, "Naive Bayes Classifier – An Ensemble Procedure for Recall and Precision Enrichment," Engineering Applications of Artificial Intelligence, vol. 136B, p. 108972, Jul. 2024, doi: 10.1016/j.engappai.2024.108972.

[39] H. J. Escalante, E. F. Morales, and L. E. Sucar, "A Naïve Bayes Baseline for Early Gesture Recognition," Pattern Recognition Letters, vol. 73, pp. 91–99, Apr. 2016, doi: 10.1016/j.patrec.2016.01.013.

[40] Y. Ma *et al.*, "The Hidden Markov Model and Its Applications in Bioinformatics Analysis," Genes & Diseases, p. 101729, Jun. 2025, doi: 10.1016/J.GENDIS.2025.101729.

[41] P. Morguet and M. Lang, "Spotting Dynamic Hand Gestures in Video Image Sequences Using Hidden Markov Models," in *Proceedings 1998 International Conference on Image Processing*, Chicago: IEEE, Oct. 1998, pp. 1–6. doi: 10.1109/ICIP.1998.999009.

[42] 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.

[43] R. W. Mellyssa, A. F. Dewi, M. Misriana, S. Suryati, and R. Rachmawati, "Pengaruh Algoritma Deep Learning dalam Meningkatkan Akurasi Sistem Pendeteksian Kondisi Jalan Raya," in *Proceeding Seminar Nasional Politeknik Negeri Lhokseumawe*, Lhokseumawe: Politeknik Negeri Lhokseumawe, Nov. 2022, pp. 12–16. Accessed: Aug. 14, 2025. [Online]. Available: https://e-jurnal.pnl.ac.id/semnaspnl/article/download/3426/2747

[44] J. S. Akosa, "Predictive Accuracy: A Misleading Performance Measure for Highly Imbalanced Data," in *Proceedings of the SAS Global Forum*, Orlando: SAS Institute Inc., Apr. 2017, pp. 1–4. Accessed: Jul. 30, 2025. [Online]. Available: https://support.sas.com/resources/papers/proceedings17/0942-2017.pdf

[45] B. Karaman, A. Güven, A. Öner, and N. S. Kahraman, "Classification of Retinitis Pigmentosa Stages Based on Machine Learning by Fusion of Image Features of VF and MfERG Maps," *Electronics (Switzerland)*, vol. 14, no. 9, p. 1867, May 2025, doi: 10.3390/electronics14091867.

[46] S. R. Bauskar, C. R. Madhavaram, E. P. Galla, J. R. Sunkara, H. K. Gollangi, and S. K. Rajaram, "Predictive Analytics for Project Risk Management Using Machine Learning," *Journal of Data Analysis and Information Processing*, vol. 12, no. 4, pp. 566–580, Sep. 2024, doi: 10.4236/JDAIP.2024.124030.

[47] J. Demšar, "Statistical Comparisons of Classifiers over Multiple Data Sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[48] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," Scientific Reports, vol. 14, no. 1, p. 6086, Mar. 2024, doi: 10.1038/s41598-024-56706-x.

[49] Suharjito, N. Thiracitta, and H. Gunawan, "SIBI Sign Language Recognition Using Convolutional Neural Network Combined with Transfer Learning and non-trainable Parameters," Procedia Computer Science, vol. 179, no. 1, pp. 72–80, Feb. 2021, doi: 10.1016/J.PROCS.2020.12.011.

Vol. 7, No. 2, April 2026, Page. 823-842
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5210

[50] P. D. Bormane and S. D. Shirbahadurkar, "Indian Sign Language Recognition: Support Vector Machine Approach," *Advances in Nonlinear Variational Inequalities*, vol. 27, no. 3, pp. 716–727, Aug. 2024, doi: 10.52783/ANVI.V27.1438.