JOURNAL OF SCIENCE
OF LAC HONG UNIVERSITY
ISSN: 2525 - 2186
Tạp chí Khoa học Lạc Hồng, **2025**, 22, 019-024

ỨNG DỤNG MEDIAPIPE VÀ SVM TRONG NHẬN DẠNG NGÔN NGỮ KÝ HIỆU HỖ TRỢ NGƯỜI KHIẾM THÍNH

Phạm Kim Don, Dương Thanh Linh*
Phân hiệu Trường Đại học Bình Dương tại Cà Mau, tỉnh Cà Mau, Việt Nam
* Tác giả liên hệ: *dtlinh.cm@bdu.edu.vn*

THÔNG TIN BÀI BÁO

Ngày nhận: 14/4/2025
Ngày hoàn thiện: 28/5/2025
Ngày chấp nhận: 28/5/2025
Ngày đăng: 15/9/2025

TỪ KHÓA

Nhận dạng ngôn ngữ ký hiệu;
MediaPipe;
Máy vector hỗ trợ;
Phát hiện điểm mốc bàn tay;
Ngôn ngữ ký hiệu Tiếng Việt.

TÓM TẮT

Ngôn ngữ ký hiệu đóng vai trò là phương tiện giao tiếp thiết yếu đối với cộng đồng người khiếm thính, tuy nhiên, việc tiếp cận và ứng dụng ngôn ngữ này vẫn gặp nhiều hạn chế tại Việt Nam do thiếu hụt nguồn lực và công cụ hỗ trợ. Nhằm góp phần khắc phục vấn đề này, nghiên cứu đề xuất một hệ thống nhận diện ngôn ngữ ký hiệu tiếng Việt dựa trên sự kết hợp giữa công nghệ MediaPipe và thuật toán phân loại SVM (Support Vector Machine). Dữ liệu huấn luyện được xây dựng từ hình ảnh các cử chỉ tay, trong đó MediaPipe đảm nhiệm việc phát hiện và trích xuất các điểm đặc trưng (landmarks) trên bàn tay. Sau đó, mô hình SVM được sử dụng để phân loại các ký hiệu. Kết quả thực nghiệm cho thấy hệ thống đạt độ chính xác từ 85% đến 90%, chứng minh tính khả thi trong việc hỗ trợ giao tiếp cho người khiếm thính thông qua công nghệ nhận dạng ngôn ngữ ký hiệu.

APPLICATION OF MEDIAPIPE AND SVM IN SIGN LANGUAGE RECOGNITION TO SUPPORT THE HEARING IMPAIRED

Pham Kim Don, Duong Thanh Linh*
Binh Duong University – Ca Mau Campus, Ca Mau Province, Vietnam
*Corresponding Author: *dtlinh.cm@bdu.edu.vn*

ARTICLE INFO

Received: Apr 14th, 2025
Revised: May 28th, 2025
Accepted: May 28th, 2025
Published: Sep 15th, 2025

KEYWORDS

Sign Language Recognition;
MediaPipe;
Support Vector Machine;
Hand Landmark Detection;
Vietnamese Sign Language.

ABSTRACT

Sign language serves as an essential means of communication for the hearing-impaired community. However, access to and adoption of sign language in Vietnam remain limited due to a lack of resources and supporting tools. To address this issue, this study proposes a Vietnamese Sign Language recognition system that combines MediaPipe technology with the Support Vector Machine (SVM) classification algorithm. The training dataset is constructed from hand gesture images, with MediaPipe responsible for detecting and extracting hand landmark features. These features are then classified using the SVM model. Experimental results demonstrate that the system achieves an accuracy rate between 85% and 90%, confirming its potential to support communication for hearing-impaired individuals through sign language recognition technology.

Doi: https://doi.org/10.61591/jslhu.22.711
Available online at: https://js.lhu.edu.vn/index.php/lachong

1. INTRODUCTION

Sign language is an important means of communication, serving as a bridge that enables the deaf community to exchange information, express thoughts, and integrate into social life. However, one of the major challenges faced by the deaf is that most people around them are not trained in sign language, leading to significant communication barriers in daily life. In Vietnam, according to data from the General Statistics Office (2023) [1], among a total of 28,887 people with disabilities in social protection centers, 4,855 individuals (accounting for 16.8%) suffer from hearing or speech impairments. Although efforts have been made to develop educational systems for the deaf, the number of people in the community who are proficient in sign language remains limited. Furthermore, there are still very few automatic systems in Vietnam that can convert sign language into text or speech, creating inconvenience in communication between the deaf and the broader community.

With the rapid development of Artificial Intelligence (AI) and Computer Vision, methods for hand gesture and sign language recognition have made significant progress. Globally, numerous studies have employed deep learning algorithms such as Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), and Transformer models to automatically recognize sign language. However, these models often require large amounts of training data and incur high computational costs, making them difficult to apply in practice especially for less common sign languages such as Vietnamese Sign Language (VSL) [2].

To address this issue, our research focuses on developing a VSL recognition system using advanced technologies such as Python, MediaPipe, OpenCV, and Scikit-learn. MediaPipe, an open-source library developed by Google, provides the capability to detect hand landmarks through its Hand Tracking model, allowing accurate spatial information to be captured for real-time interpretation of hand signs. OpenCV supports efficient image processing, enabling the system to operate smoothly with live video streams. The system applies the Support Vector Machine (SVM) classification algorithm from Scikit-learn, a powerful tool that enables high-accuracy hand sign recognition even with limited training data. SVM allows the system to be flexible and effective in recognizing diverse sign styles and various hand movements. Our objective is to develop a low-cost, computationally efficient sign language recognition solution that can be widely applied in real-world scenarios, thereby contributing to bridging the communication gap between the deaf community and the broader society.

The remainder of this paper is organized as follows: Section 2 presents related work. Section 3 introduces the methodologies, including MediaPipe and SVM. Section 4 describes the system architecture. Section 5 discusses experimental results, and Section 6 concludes the study with future directions.

2. RELATED WORK

Sign Language Recognition (SLR) is an important research field aimed at supporting the deaf community in communicating with society. In recent years, many studies have applied artificial intelligence to SLR, especially deep learning models such as CNN, LSTM, and YOLO. Zhang et al. (2021) [3] employed a combination of CNN and LSTM to recognize American Sign Language (ASL) with an accuracy of 95%, while Kumar et al. (2020) [4] applied YOLO for real-time gesture recognition with high performance. However, these methods require significant computational resources, making deployment on mobile devices challenging. MediaPipe is a powerful tool for extracting hand features without the need for specialized hardware. Zhang et al. (2022) [5] combined MediaPipe with CNN to recognize ASL, achieving an accuracy of 97%, while Abdul et al. (2021) [6] used MediaPipe with KNN to recognize Indian Sign Language (ISL) with an accuracy of 92%. However, deep learning models still require large amounts of training data to achieve optimal performance. SVM is a widely used classification algorithm in SLR due to its ability to classify accurately with low computational cost. Rahman et al. (2019) [7] applied SVM on Leap Motion data to recognize British Sign Language (BSL), achieving 92% accuracy. In another study, Nguyen et al. (2021) [8] integrated MediaPipe with SVM for gesture recognition in educational applications, demonstrating high efficiency even with small datasets. In the context of VSL, the number of studies remains limited. Le et al. (2020) [9] used CNN to recognize the VSL alphabet but faced challenges related to data availability. Hoang et al. (2021) [10] applied RNNs but encountered high computational costs. These limitations highlight the need for a more optimal approach to recognizing VSL one that achieves high accuracy, low cost, and ease of deployment.

Building upon previous studies, this research proposes a method for VSL recognition by combining MediaPipe and SVM. The use of MediaPipe allows for fast and accurate hand feature extraction, while the SVM classifier can effectively recognize gestures without requiring a large training dataset.

3. RESEARCH METHODOLOGY

3.1 MediaPipe

MediaPipe is a cross-platform, open-source framework developed by Google that enables the construction of data perception pipelines from various input signals, including video and audio. Designed to optimize real-time perceptual data processing, MediaPipe stands out for its ability to efficiently deploy processing modules on mobile platforms and systems with limited computational resources. Compared to other technologies such as OpenPose, MediaPipe offers superior advantages, particularly in terms of accuracy and computational performance [11]. One of MediaPipe's key strengths lies in its use of regression analysis to determine finger coordinates, which helps narrow the detection range while maintaining high accuracy. MediaPipe extracts 21 three-dimensional (3D)

coordinates from a single camera frame, with the X and Y coordinates normalized based on the bounding box. The Z coordinate, representing depth, plays a crucial role in determining the distance between objects and accurately assessing the position of hands and fingers in 3D space [12]. This approach ensures advanced hand and finger tracking with minimal processing requirements, making it particularly suitable for low-performance environments such as mobile platforms. These capabilities highlight MediaPipe's effectiveness compared to other models.

3.2 Hand Landmark model

The Hand Landmark model in MediaPipe is designed to accurately detect and locate 21 three-dimensional (3D) landmarks on a hand from a single input image. This process consists of two main steps: First, a palm detection module identifies the location of the hand in the image. This detector operates on the entire image and returns an oriented bounding box that defines the region containing the hand. Once the hand region is localized, the second model precisely estimates the 21 3D landmarks of the finger joints within this region using a direct regression method. These landmarks include fingertips, finger joints, and the palm base.

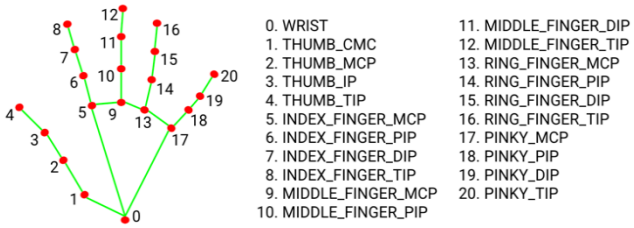

Figure 1. *Hand LandMarks*

The use of MediaPipe's Hand Landmark model in this study enables accurate identification of the positions and movements of fingers, forming the foundation for recognizing hand gestures in VSL. Figure 1 illustrates the landmarks detected by MediaPipe on the hand in various poses. These landmarks represent finger joints and knuckles, which are connected to form a complete hand structure. This structure allows for effective feature extraction to support the gesture classification process.

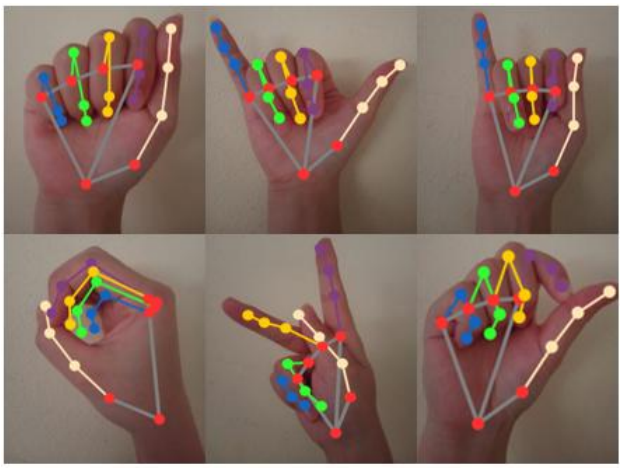

Figure 2. *Sample Hand Gestures*

Figure 2 shows examples of various hand gestures used in Vietnamese Sign Language, with 3D hand landmarks detected by MediaPipe. Each gesture is represented by 21 key points that form the finger joints and palm structure. The landmarks are color-coded by finger and connected to visualize the orientation and posture of the hand. These gesture patterns serve as input data for the classification model in the recognition system.

4. SYSTEM ARCHITECTURE

The system architecture consists of three main phases. In stage 1, input frames are processed using MediaPipe to detect hands and extract 21 landmarks, which are saved in CSV format. In stage 2, the collected data undergoes pre-processing, including data cleaning, normalization, and division into training and validation sets.

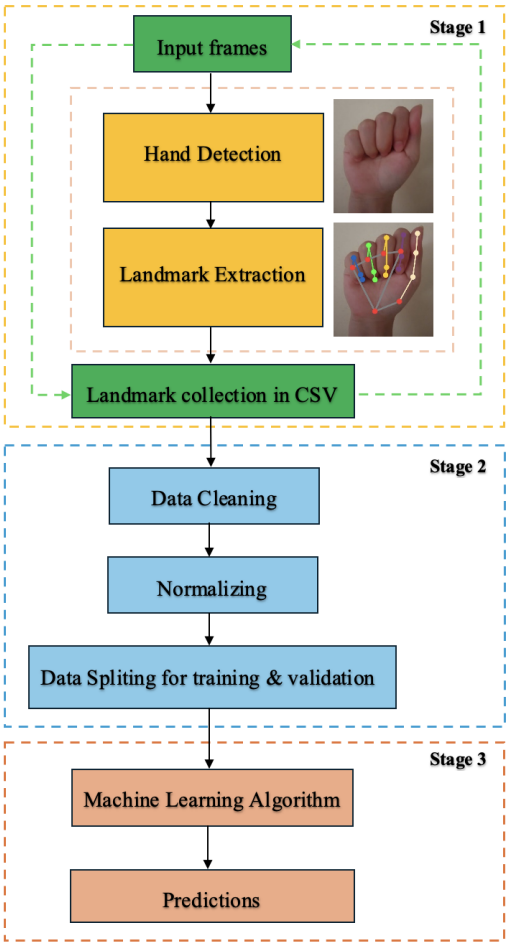

Figure 3. *System Architecture for Hand Gesture Detection and Sign Language Recognition*

In stage 3, a machine learning algorithm is applied to the processed data to perform gesture classification and generate predictions. This process enables real-time sign language recognition with high accuracy and efficiency.

4.1 Stage 1: Hand Image Collection and Feature Extraction

The first stage focuses on collecting and processing input image data to prepare for the machine learning model. The system begins by receiving input frames from a camera or

video stream. Then, the MediaPipe algorithm is applied to detect the hand and accurately identify its location within the frame. Next, MediaPipe extracts 21 hand landmarks representing finger joints and positions. These landmark coordinates are stored in CSV files, creating a structured dataset for the subsequent data processing steps.

4.2 Stage 2: Data Processing, Cleaning, and Normalization Before Training

After collecting the raw data, the system enters the data processing phase to ensure the quality and effectiveness of the machine learning model. First, the data undergoes a cleaning process to remove erroneous or incomplete records, such as frames where no hand was detected or data affected by unstable lighting conditions or camera angles. After cleaning, the data is normalized to bring all values into a consistent range, minimizing the impact of factors like hand size and camera distance, thereby improving the model's learning capability. Finally, the dataset is split into 80% for training and 20% for testing, ensuring that the model learns meaningful patterns and has sufficient data to evaluate its performance.

4.3 Applying Machine Learning Algorithm and Gesture Prediction

In stage 3, the SVM classification algorithm is applied to the processed data to perform gesture recognition and generate predictions. SVM is selected due to its effectiveness in high-dimensional data spaces and its robustness with limited training data.

The SVM model was chosen for its effective classification capabilities in high-dimensional spaces and its flexibility when working with large datasets. Specifically, the Radial Basis Function (RBF) kernel is employed to transform the data from the input space to a higher-dimensional feature space, enabling better separation of non-linear data classes. The formula applied in this study illustrates the optimization process of SVM as follows:

$$f(x) = \text{sign}\left(\sum_{i=1}^{N} \alpha_i y_i K(x_i, x) + b\right) \quad (1)$$

In Equation (1), x is the new data point to be classified *(the extracted features from a hand gesture image)*, and x_i are the data points in the training set. Each training point x_i is associated with a label y_i, representing the class it belongs to (gestures "A", "B", etc.). The coefficients α_i are learned during the training process, indicating how much influence each training point has on the classification decision. The kernel function $K(x_i, x)$, with RBF defined as $exp(-\gamma|x - x_i|^2)$, measures the similarity between the new data point and the training points. Finally, b is the bias term used to adjust the decision boundary between the classes.

The classification process is based on computing the weighted sum of the influences from the training points, followed by applying the sign function to determine which class the new point belongs to. The use of the RBF kernel in this study enables the model to flexibly separate non-linear data classes in the feature space, thereby improving the accuracy of hand gesture recognition. This approach is well-suited to the complex and diverse nature of hand shapes in sign language [13].

To ensure the quality and evaluate the performance of the model, hand gesture features were extracted using MediaPipe, including the coordinates of hand landmarks and geometric parameters such as the distances between finger joints (Figure 1). After training, the model was evaluated using a Confusion Matrix to determine the metrics: True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN) [14]. The key evaluation metrics include:

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

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

$$F1 = 2 * \frac{P \cdot R}{P + R} \quad (4)$$

Precision (2) measures the accuracy of the model in correctly classifying positive instances, while Recall (3) reflects the model's ability to identify all actual positive instances. The F1-score (4) represents the harmonic mean of Precision and Recall, providing a balanced evaluation of both metrics. The application of these formulas enables a more detailed performance analysis of the model, allowing comparisons across different model versions when tuning parameters or modifying the training dataset. Consequently, the study can determine appropriate optimization strategies, ranging from enhancing the quality of input features to refining the training algorithm in order to achieve higher classification accuracy.

5. EXPERIMENTAL SETUP RESULTS

5.1 Data Collection

The data samples were collected through a web-based application developed using Python. The web platform supports functions such as image capturing, image processing, and recording hand tracking landmarks as feature extraction results. The images were captured using a standard webcam integrated with a laptop (figure 4).

Figure 4. *Sample image captured directly from the webcam during data collection*

The hand gestures correspond to 23 letters of the VSL alphabet, with 100 images captured for each gesture. Images from the webcam are displayed directly on the screen, allowing participants to monitor and verify the data collection process in real time.

A total of 2,400 images were collected directly, with each hand gesture represented by a dataset of 100 images. The dataset includes Vietnamese sign language symbols: "a, b, c, d, đ, e, g, h, i, k, l, m, n, o, p, q, r, s, t, u, v, x, y". The data was then split in an 80:20 ratio for training and testing. This approach ensures that the model has sufficient data to learn and improve its predictive capabilities, while also maintaining an independent test set to evaluate the model's accuracy and generalization on unseen data.

5.2 Classification Results and Performance Evaluation

In the evaluation of the Vietnamese sign language recognition model's performance, the analysis of the confusion matrix plays a crucial role in determining the model's accuracy and overall effectiveness. The results indicate that the model performs well, with most predictions concentrated along the main diagonal of the matrix, demonstrating its ability to accurately recognize the majority of signs. This concentration not only reflects the efficiency of the model but also highlights the thorough preparation of the training process, which utilized appropriate data to achieve high performance.

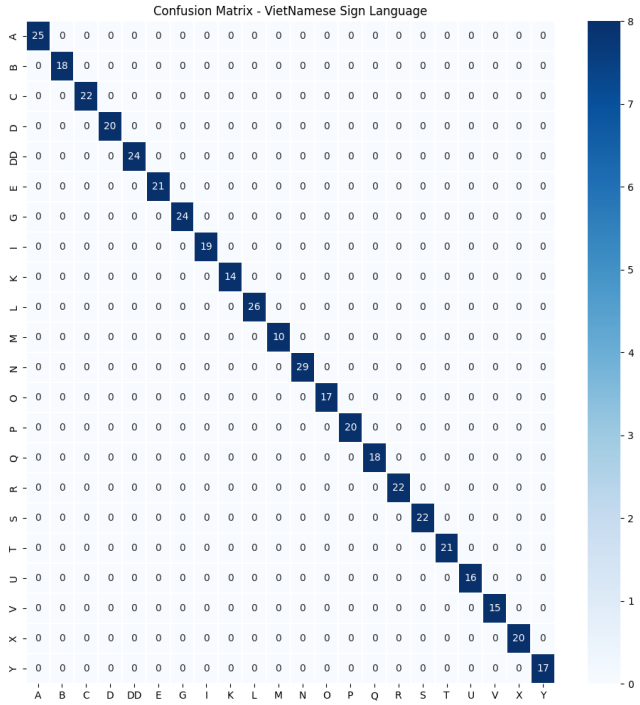

Figure 5. *Experimental Data Represented in the Confusion Matrix. This figure illustrates the model's performance on the test dataset using a confusion matrix. The majority of predictions are concentrated along the main diagonal, demonstrating high classification accuracy across most sign classes. The absence of values in the off-diagonal cells further indicates the model's robustness and minimal misclassification rate, even in the presence of class imbalance.*

However, alongside these positive indicators, a deeper analysis of the misclassification errors (i.e., the off-diagonal cells in the confusion matrix) is equally essential. These errors may reveal specific signs that the model struggles to distinguish. Identifying frequently confused sign pairs can provide insight into visual or motion-related similarities between them, which may contribute to the model's classification challenges.

The data reveals a significant class imbalance in the number of samples. Specifically, class M contains only 10 samples, which is notably fewer than class N with 29 samples. Similarly, classes K (14 samples), O (17 samples), U (16 samples), and V (15 samples) also have substantially fewer samples compared to other classes. However, no samples were misclassified as belonging to another class (all off-diagonal cells have a value of zero), indicating that the model performs exceptionally well.

Finally, to achieve a more comprehensive evaluation, the metrics of Precision, Recall, and F1-score can be calculated for each individual class. However, in the case of no misclassifications, as illustrated in Figure 5, these metrics all reach their maximum value of 1.0 (or 100%), indicating perfect accuracy of the model on the current test dataset. The model demonstrates highly promising performance metrics. An accuracy of 97.50% reflects strong classification capability, while a misclassification rate of only 2.50% indicates minimal prediction errors. In addition, a loss value of 0.03 suggests that the model is effectively learning and producing predictions that closely align with actual values. Based on these outstanding evaluation metrics, the model shows great potential for practical applications. Achieving such a high level of accuracy implies strong generalizability and suitability for deployment across various systems. Moreover, the low error rate and negligible loss further highlight the algorithm's efficiency, minimizing mistakes during the prediction process.

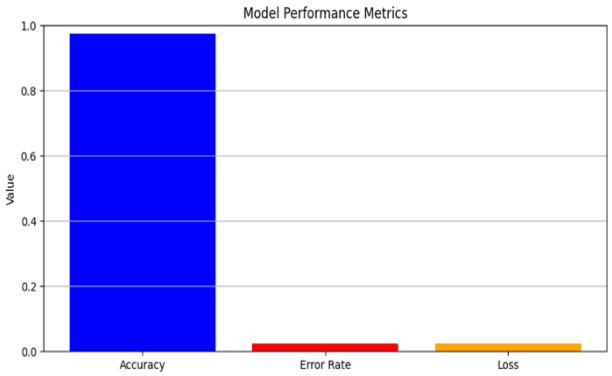

Figure 6. *Provides a visual representation of key model evaluation metrics, including Accuracy, Error Rate, and Loss.*

Figure 7. *Real-time Vietnamese Sign Language Recognition.*

As illustrated in Figure 6, the evaluation metrics including accuracy, error rate, and loss show that the model performs with high reliability. Furthermore, Figure 7 demonstrates the system's ability to recognize Vietnamese Sign Language in real-time, as shown with the example letter sequence "X I N C H A O" meaning "Hello".

6. CONCLUSION

This study proposed a VSL recognition system using MediaPipe and SVM, which achieved high accuracy, demonstrating both feasibility and effectiveness. The use of MediaPipe for hand feature extraction significantly reduces hardware requirements, while SVM ensures precise classification even with small datasets. However, to enhance generalizability, future improvements should include dataset expansion and algorithm optimization. Looking ahead, the system can be integrated into mobile applications to support communication for the hearing impaired, contributing to a more inclusive society.

7. REFERENCES

[1] Tổng cục Thống kê Việt Nam. "*Báo cáo Điều tra người khuyết tật năm 2023*." Bộ Kế hoạch và Đầu tư. Hà Nội, Việt Nam.

[2] Zhang, X., Liu, Y., & Wu, J. "Sign Language Recognition: A Comprehensive Review." *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 10, pp. 4302-4320, **2021**.

[3] Zhang, Y., Wang, L., & Chen, H. (**2021**). Sign language recognition based on convolutional neural network and LSTM. *Multimedia Tools and Applications*, 80(5), 7051–7067.

[4] Kumar, P., Sharma, K., & Kaushik, S. (**2020**). A deep learning-based framework for real-time sign language recognition using YOLO. *Pattern Recognition Letters*, 140, 27–34.

[5] Zhang, L., Li, X., & Ma, J. (**2022**). Real-time sign language recognition using MediaPipe hand tracking and convolutional neural networks. *Journal of Visual Communication and Image Representation*, 84, 103478.

[6] Abdul, R., Khan, M., & Iqbal, S. (**2021**). Indian sign language recognition using MediaPipe and KNN. *International Journal of Advanced Computer Science and Applications*, 12(6), 89–95.

[7] Rahman, F., Hossain, M., & Kim, J. (**2019**). Hand gesture recognition for sign language using support vector machine with optimal feature selection. *Journal of Ambie Intelligence and Humanized Computing*, 10(11), 4383–4394.

[8] Nguyen, T. D., Pham, V. H., & Tran, M. Q. (**2021**). Integrating MediaPipe and SVM for sign language recognition in educational applications. *Vietnam Journal of Computer Science*, 8(3), 123–135.

[9] Le, H. T., Nguyen, D. K., & Pham, A. T. (**2020**). Vietnamese sign language recognition using deep learning. *Proceedings of the International Conference on Artificial Intelligence and Data Science*, 124–129.

[10] Hoang, T. P., Bui, N. H., & Tran, T. D. (**2021**). A recurrent neural network approach for Vietnamese sign language recognition. *Neural Computing and Applications*, 33, 10567–10579.

[11] Rane, A., Bhosale, D., Jadhav, S., & Lande, S. (**2023**). Sign Language Detection of English Alphabets for Deaf and Dumb People. *International Journal for Research in Applied Science and Engineering Technology (IJRASET)*, 11(III), 1743–1747.

[12] Google. (n.d.). **MediaPipe Hands**. Retrieved February 8, **2025**, from https://mediapipe.readthedocs.io/en/latest/solutions/hands.html

[13] Schölkopf, B., & Smola, A. J. (**2002**). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press.

[14] Han, J., Kamber, M., & Pei, J. (**2011**). *Data Mining: Concepts and Techniques* (3rd ed.). Elsevier.