Multimodal AI Framework for Sign Language Recognition and Medical Informatics in Hearing-Impaired Patients

Pratya Nuankaew[1,*], Parin Khamthep[2], Patdanai Jaitem [3], Kuljira S. Nuankaew [4], Kaewpanya S. Nuankaew[5], Wongpanya S. Nuankaew[6]

[1]*Department of Digital Business, School of Information and Communication Technology, University of Phayao, Phayao 56000, Thailand*

[2,3,6]*Department of Computer Science, School of Information and Communication Technology, University of Phayao, Phayao 56000, Thailand*

[4,5]*High School Student, Phayao Pittayakhom School, Phayao 56000, Thailand*

(Received: July 25, 2025; Revised: September 20, 2025; Accepted: December 18, 2025; Available online: January 14, 2026)

Abstract

This study assesses the feasibility of YOLO-based detectors for recognizing Thai Sign Language (TSL) in clinical intake workflows. This study evaluates four YOLO variants (YOLOv8, YOLOv9, YOLOv11, and YOLOv12) on a sign-language–based clinical dataset using Precision, Recall, mAP@50, and mAP@50:95. YOLOv11 achieves the strongest overall performance, attaining Precision of 0.9527, Recall of 0.9700, mAP@50 of 0.9723, and mAP@50:95 of 0.6679, while maintaining a compact model size (7.8 MB) and moderate inference latency (16.1 ms). In comparison, YOLOv8 and YOLOv9 provide faster inference but lower accuracy, whereas YOLOv12 exhibits reduced detection performance despite increased model size and latency. Experiments were conducted on a dataset of 10,956 images spanning 95 classes under realistic clinical recording conditions, including variations in viewpoint, illumination, motion, and partial occlusion. The results demonstrate the effectiveness of YOLO-based models for visual sign interpretation while highlighting ongoing challenges in fine-grained localization and robust generalization in real-world clinical environments. These findings support a multimodal pipeline that uses an image-based detector as the core perception component, supplemented with pose/key point cues, OCR, and NLP layers to convert recognized signs into structured medical intents for triage and telemedicine. Future work will focus on sequence-level evaluation, expanding dialectal and co-articulated TSL coverage, and developing compression or distillation techniques to enable reliable on-device inference in resource-limited settings.

Keywords: Continuous Sign Language Recognition, YOLO-based Detection, Thai Sign Language, Pose Estimation, Medical Informatics

1. Introduction

Communication barriers significantly hinder healthcare for the Deaf and Hard-of-Hearing (DHH). In clinical settings, deaf patients rely on sign language to communicate symptoms and understand medical advice, but professional interpreters are often unavailable. This gap leads to misdiagnoses and decreased access to care: for example, hospitals without sign interpreters make it difficult for deaf patients to express physical complaints or follow treatment plans [1]. Indeed, studies highlight that Deaf individuals who cannot understand spoken language face unequal access to healthcare services [2]. These challenges emphasize the need for automated sign language recognition and translation systems. Recent advances in deep learning have demonstrated strong performance in sign-gesture recognition, but most studies have been conducted under controlled laboratory conditions and on limited vocabularies. For example, Alsharif et al. reported 98.2% mAP using a YOLOv11 model combined with MediaPipe hand tracking for the recognition of 26 isolated ASL alphabet letters on a dataset of approximately 34,000 images collected in a non-clinical environment [3], [4]. Although the result indicates the potential of modern detectors, the study did not involve continuous signing, domain-specific medical vocabulary, or diverse signer profiles, and therefore cannot be assumed to generalize to real healthcare settings where lighting, motion, and communication context vary widely.

Modern sign language recognition commonly applies deep neural networks to capture both spatial and temporal characteristics of gestures. Deep neural networks such as combinations of CNN and LSTM have shown promising

***Corresponding author: Pratya Nuankaew (nuankaew.p@gmail.com)**
DOI: https://doi.org/10.47738/jads.v7i1.1096
This is an open access article under the CC-BY license (https://creativecommons.org/licenses/by/4.0/).
© Authors retain all copyrights

results in capturing both spatial and temporal aspects of sign language gestures, challenges remain in adapting these models to real-world conditions including motion blur, occlusion, and signer variation, especially in clinical environments where lighting and backgrounds are uncontrolled [5], [6], [7]. Previous studies combining visual inputs with skeletal keypoint streams have reported high accuracy in isolated sign classification tasks [7]. However, these results are typically achieved in controlled environments with limited signer variability, stable lighting, and static backgrounds. In real deployment—especially in clinical settings—performance can degrade due to motion blur, occlusion, differences in signing speed and style, and the need to recognize continuous rather than single-frame gestures. These constraints highlight the gap between laboratory performance and real-world applicability, reinforcing the need for more robust multimodal models.

In healthcare informatics, integrating diverse data modalities has become a powerful strategy. Multimodal AI frameworks combine inputs such as clinical text, physiological signals, images, and structured electronic health records have shown better prediction and decision-support performance than single-modal approaches. This evidence reinforces the value of unified systems that combine gesture recognition with comprehensive healthcare data [8], [9], [10]. For example, the HAIM framework showed that models using tabular data, time series, text, and images together outperformed single-modality models (improving diagnostic accuracy by 6–33% in chest X-ray tasks) [8]. Overall, AI systems that leverage multiple data sources tend to provide more accurate and reliable results in medical applications [8]. Similarly, a clinical sign-language tool should combine visual gesture recognition with patient-specific medical context. A multimodal approach for deaf patient care might merge sign-to-text translation with electronic health record (EHR) data or symptom lexicons, creating a comprehensive view of the patient's history in real time. Such a unified model could process video of a patient signing, analyze the sequence with CNN-LSTM networks, and generate medical terms linked to the patient's record. This approach contrasts with previous systems that handle sign translation in isolation, and it aims to assist history-taking by integrating translations directly into medical informatics workflows [8].

In the Thai context, these needs are urgent. Thai Sign Language (TSL) is the official language for deaf communities, yet Thai medical facilities rarely have interpreters available. Some prototype solutions exist, such as a recent mobile app that translates a set of 60 health-related Thai signs using on-device deep learning [11]. While promising, these tools only cover limited vocabulary and simple queries. Academic research on Thai sign recognition, like for digits and the alphabet, shows that it is feasible [12], but also suggests that more vocabulary and optimization are necessary before it can be used in healthcare. No existing system fully integrates Thai sign interpretation with clinical data entry or dialogue.

To address these gaps, researchers propose a Multimodal AI Framework for Sign Language Recognition and Medical Informatics tailored for Thai hearing-impaired patients. Our main contribution is an end-to-end model that recognizes Thai sign language sequences using deep learning and connects the output to medical informatics tasks. The core of the model is a CNN-based vision encoder combined with a temporal sequence model, such as BiLSTM or Transformer, that processes sign language videos. By utilizing skeleton key points via MediaPipe or similar tools and CNN feature maps, the model captures detailed hand and body movements [7]. The recognized sign language is then mapped to medical terminology and integrated with patient context, enabling automated history-taking and record updates. In this way, our framework goes beyond isolated gesture recognition: it combines sign translation with healthcare data to support real-time, accessible consultations. This work advances the state of the art by focusing on model development rather than, for example, chatbots, and by situating it within Thailand's healthcare system. Ultimately, it aims to empower deaf patients and clinicians alike by reducing communication barriers and enhancing the inclusivity of medical informatics.

This study aims to develop a multimodal AI model for Thai Sign Language recognition and translation using advanced deep learning to achieve high accuracy and practical utility in healthcare. It also seeks to design and implement a framework that integrates sign language recognition with medical history collection for hearing-impaired patients, routing outputs into medical informatics systems to support systematic documentation and clinical data management. Additionally, the project aims to build a prototype capable of processing Thai Sign Language videos and translating them into medically relevant text in real time, suitable for deployment in Thai hospitals and clinics.

Therefore, the main goal of this research is to create a multimodal artificial intelligence model capable of accurately recognizing and translating Thai sign language in medical settings. It also strives to develop a framework that combines sign language recognition with medical informatics to improve efficient history-taking for hearing-impaired patients. The ultimate aim is to reduce communication barriers in Thai healthcare facilities, promote equal access to medical services, and lay the groundwork for future digital health tools tailored to the Thai sociocultural and clinical context.

2. Literature Review

2.1. AI for Sign Language in Healthcare

In contemporary research, AI-based sign language recognition has advanced rapidly. A recent survey highlights that multimodal neural network models combining vision and sensor inputs achieve much higher accuracy than unimodal approaches [13]. Contemporary deep-learning systems have evolved from recognizing isolated static gestures to translating continuous sign-language streams in nearly real time [13]. However, existing algorithms still lack the robustness and generalization necessary for widespread commercial use [13]. For example, Aly and Fathi's hybrid Transformer CNN model reported a remarkably high accuracy of 99.97% on static ASL gesture recognition [14], showcasing the effectiveness of recent architectures in sign recognition tasks.

A major distinction in sign language research lies between isolated gesture recognition and continuous sign language recognition. Isolated models classify single signs or alphabets from short clips or individual frames, often achieving high accuracy because there is no temporal dependency between gestures. In contrast, continuous sign recognition requires understanding co-articulation, transition movements, signer-specific variation, and semantic dependencies across sequences, making it a significantly more complex task that typically demands temporal models such as Transformers, CTC-based architectures, or sequence-to-sequence decoders. Most high-accuracy studies in the literature, including YOLO-based methods, focus on isolated recognition, which limits their applicability in real medical consultations where patients communicate in full sentences rather than discrete tokens.

Beyond isolated gesture recognition, recent advances in continuous sign language recognition (CSLR) have focused on sequence-level modeling using Transformer architectures, Connectionist Temporal Classification (CTC), and encoder–decoder frameworks. Unlike frame-based classification, CSLR requires temporal alignment between video frames and linguistic units, handling co-articulation, motion continuity, and signer variability. State-of-the-art models such as the Sign Language Transformer and VAC (Visual Alignment Constraint) have demonstrated significant improvements on large-scale datasets such as PHOENIX-2014 and CSLR benchmarks, outperforming traditional CNN-LSTM pipelines. These works highlight that continuous recognition is essential for real clinical communication, where patients narrate symptoms in full sentences rather than isolated signs [11].

2.2. Medical Sign Language Translation

In the medical context, preliminary studies are creating specialized sign-language translation systems. For example, Roelofsen et al. (2021) developed a system that translates common COVID-19 healthcare phrases into Dutch Sign Language (NGT) using video and avatar animations [15]. This method shows architecture that could be expanded to other medical terms and sign languages. More broadly, sign-language interpretation research integrates features from hand gestures, facial expressions, and lip movements to translate signs into text or speech [16]. Such systems help bridge the communication gap between deaf patients and healthcare providers by presenting medical instructions in accessible formats [17].

2.3. AI and Medical Informatics for the Hearing-Impaired

In medical informatics, AI is used to improve healthcare access for the deaf and hard of hearing. For example, AI-powered speech-to-text algorithms now offer live captioning during telemedicine sessions or in-person appointments [18]. In audiology, AI-driven tools automate hearing assessments and tailor device fittings, such as hearing aids and cochlear implants, thereby aiming to boost clinicians' efficiency and provide more personalized patient care [19]. These AI applications help streamline communication and clinical workflows, making healthcare information and services more inclusive for those with hearing impairments.

While the HAIM framework demonstrates multimodal integration for medical decision support, our study adapts the same principle in the context of sign-language based clinical communication. Instead of combining radiology images, tabular vitals, and clinical notes, the proposed framework fuses visual gesture detection, skeletal key points, and medical intent mapping, and routes the output into structured EHR fields. Thus, the present model operationalizes HAIM's core idea—linking heterogeneous modalities into a unified clinical workflow—but focuses on patient–provider communication rather than diagnostic prediction.

In summary, the reviewed literature shows notable progress in AI-driven sign language recognition, especially with deep learning architectures that reach high accuracy in gesture classification. However, research on medical sign language translation remains limited, often focusing on specific vocabulary or narrow contexts. Moreover, the integration of sign language recognition outputs into medical informatics systems is still underdeveloped, particularly within the Thai healthcare system. In this context, the present study aims to create a multimodal AI framework that not only recognizes and translates Thai sign language but also integrates smoothly with medical history-taking and electronic health records. This research aims to reduce communication barriers in clinical practice and promote equitable access to healthcare services for hearing-impaired patients in Thailand.

3. Materials & Methods

3.1. Research Design and Framework

The conceptual framework diagram illustrates the Multimodal AI Framework for Sign Language Recognition and Medical Informatics in Hearing-Impaired Patients, which consists of three main parts: inputs, processing model, and outputs (see figure 1). The inputs include sign language videos and medical context data, both of which are processed by the Multimodal AI Model. This model translates sign language into text and assists with medical history-taking. The resulting information is then stored in the Electronic Health Record (EHR) system, ensuring accurate clinical documentation and promoting equitable healthcare access for hearing-impaired patients.

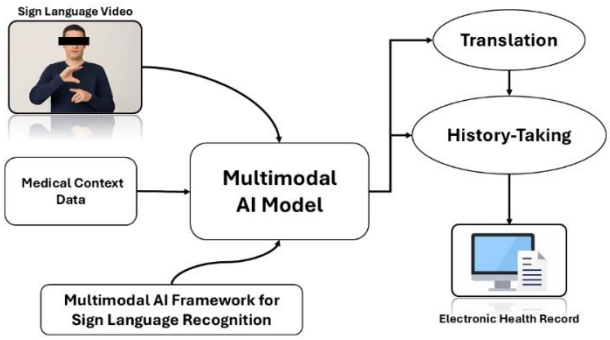

Figure 1. Research Design and Framework

3.2. Dataset Collection and Preparation

3.2.1. Sign Language Dataset.

The selection of 60 symptom categories and 35 history-taking question categories was based on clinical relevance rather than arbitrary grouping. The initial taxonomy was derived from the standard outpatient history-taking form used in Thai Ministry of Public Health hospitals, following the SOAP structure (Subjective, Objective, Assessment, Plan) and aligned with WHO guidelines for patient intake. The proposed categories were reviewed by two licensed physicians and one registered nurse to ensure that they accurately reflect real medical interviews and symptom-reporting workflows. In addition, a certified Thai Sign Language (TSL) interpreter and a Deaf consultant validated the linguistic suitability of each category to confirm that the signs correspond to expressions naturally used by TSL users in medical communication. This validation process ensured that the dataset is both clinically meaningful and linguistically grounded, supporting real-world applicability rather than laboratory-only relevance.

 ISSN 2723-6471

The source videos have been converted into images with a resolution of 640×640 pixels to support the training of detection and recognition models. A table shows the number of images per class for training, validation, and testing, as presented in table 1. These include various clinical data such as address, age, blood pressure, BMI, temperature, and others. Examples of data collected during the process are illustrated in figures 2 and 3.

Table 1. Sign Language Dataset

Class	Training	Validation	Test	Total
Address	215	54	54	323
Age	209	53	53	315
Gender	214	54	54	322
Blood Pressure	210	53	53	316
Weight	211	53	53	317
Height	210	53	53	316
BMI	207	52	52	311
Temperature	208	52	52	312
Heart Rate	211	53	53	317
Allergy	214	54	54	322
Chronic Disease	210	53	53	316
Family History	213	54	54	321
Current Medication	209	53	53	315
Previous Surgery	208	52	52	312
Smoking Habit	211	53	53	317
Alcohol Consumption	210	53	53	316
Exercise Habit	214	54	54	322
Sleep Pattern	207	52	52	311
Stress Level	208	52	52	312
Occupation	211	53	53	317
Marital Status	210	53	53	316
Education Level	213	54	54	321
Nationality	209	53	53	315
Religion	208	52	52	312
Emergency Contact	211	53	53	317
Health Insurance	210	53	53	316
Chief Complaint	213	54	54	321
Present Illness History	209	53	53	315
Past Illness History	208	52	52	312
Medication and Treatment History	211	53	53	317
Family Health History	210	53	53	316
Lifestyle and Risk Factors	213	54	54	321

ISSN 2723-6471

Class	Training	Validation	Test	Total
Address	215	54	54	323
Travel History	209	53	53	315
Immunization History	208	52	52	312
Diet/Nutrition	211	53	53	317
Total	7,286	1,835	1,835	10,956

An analysis of the dataset revealed an observable class imbalance across the 95 categories, where certain medical history classes (e.g., Exercise Habit, Family Health History) contain more than 320 samples, while others have fewer than 300 samples. Although the dataset is not severely skewed, this imbalance may affect model learning by biasing predictions toward majority classes.

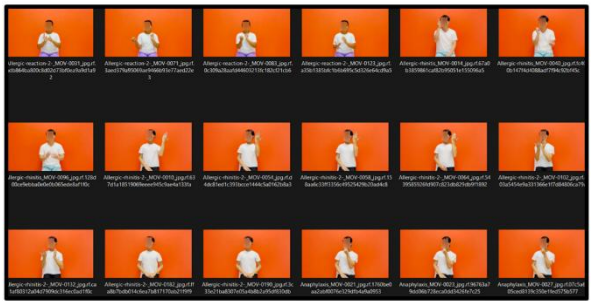

Figure 2. Examples of data collected

Figure 3. Examples of data collected

Figure 2 and figure 3 present example images from the dataset, including both male and female subjects performing different gesture- and symptom-related actions. The samples were collected at varying times and with different background colors, demonstrating diversity in subjects, poses, and acquisition conditions for model training and evaluation.

3.2.2. Annotation Process

The annotation process was carried out by three annotators: one certified Thai Sign Language interpreter and two trained research assistants. A detailed labeling protocol was developed to ensure consistency in bounding-box placement and class assignment, including rules for hand visibility, partial occlusion, dominant-hand priority, and minimum box coverage. To assess label reliability, 15% of the dataset was double-annotated, and inter-annotator agreement was measured using Cohen's Kappa [20], [21], yielding $\kappa = 0.87$ for class labels and $\kappa = 0.82$ for bounding-box placement, indicating strong agreement. Discrepancies were reviewed in weekly adjudication sessions led by a senior annotator, and final consensus labels were stored as ground truth. Additional quality control was performed through random sampling and visual inspection before model training to prevent mislabeled or ambiguous samples from entering the dataset.

3.2.3. Data Augmentation

To enhance robustness, the training set will undergo data augmentation, including rotation, horizontal and vertical flipping, brightness and contrast adjustments, random cropping and scaling, and mild blurring [22]. Bounding box and keypoint coordinates will be transformed consistently using augmentation utilities that naturally handle detection and pose labels [23]. This process aims to reduce overfitting and improve generalization under varying lighting and camera angles typical of Thai clinical environments.

To evaluate the effect of augmentation on model robustness, a controlled ablation experiment was conducted by training YOLOv9 with and without augmentation under the same hyperparameters. The augmented model achieved a +2.8% improvement in overall mAP@50 and a +4.2% improvement in mAP@50:95, with the largest gains appearing in classes recorded under low-light and side-view camera angles. Recall for motion-influenced signs (e.g., Exercise

Habit, Past Illness History) increased from 0.91 to 0.95, suggesting that motion-based augmentations helped the model generalize better to real clinical video conditions.

3.3. Multimodal Feature Extraction

3.3.1. Visual Features

The image stream served as the main source of features. The researchers used YOLOv11 to detect and classify Thai sign language gestures in real time from video images, focusing on two main categories of the project: 60 symptom clusters and 35 history-taking questions. Before choosing a model, YOLO versions 8 through 12 were compared for performance specifically on sign language interpretation and history-taking tasks.

3.3.2. Skeletal Features

To address variations in signer pose, researchers have adopted multi-stream frameworks that integrate skeletal and visual information. Pose estimation techniques are used to extract hand and body keypoints as skeletal streams, while CNN- or YOLO-based models provide complementary visual features. By combining these streams, the approach captures both the spatial configuration of articulated joints and detailed appearance cues. This reflects the understanding that sign language interpretation involves more than hand shape alone, relying also on hand position, body posture, and movement patterns [4], [24]. Previous studies have shown that merging skeletal and visual features within deep learning systems leads to more accurate and reliable sign language recognition.

3.4. Model Architecture and Training

3.4.1. Deep Learning Backbone.

The experimental setup initially included YOLOv8 through YOLOv12 for comparative analysis. The execution environments were Python on Google Colab with GPU/TPU acceleration and PyTorch/TensorFlow, including preprocessing and data validation for reproducibility.

3.4.2. Training Process

To address the effects of class imbalance during model training, class weights were incorporated into the loss function to give more penalty to underrepresented classes. This method was chosen over oversampling because oversampling can create redundancy and increase the risk of overfitting, especially in gesture-based datasets where samples tend to be visually similar. Using class-weighted loss has been shown to enhance convergence stability in object detection models without expanding dataset size or training time, making it appropriate for this use case.

Training used a 640×640 input as specified in the data.yaml, with Automatic Mixed Precision (AMP) enabled and a log period set to 10, which involved performing checkpoints every 10 epochs. This process was carried out over 100 to 150 epochs with a dataset size of 32. During training, the box loss, classification, and objectness (or DFL) were monitored, and various performance metrics, including Precision, Recall, mAP@50, and mAP@50–95 [25], were reported. The researchers followed YOLO's default and scheduled learning rate policy during the training phase.

3.4.3. Model Selection Rationale

This study evaluates YOLOv8, YOLOv9, YOLOv11, and YOLOv12 based on both reported accuracy and architectural innovations relevant to deployment on resource-limited devices. Each model introduces design refinements targeting improved feature representation, computational efficiency, and training stability. YOLOv8 adopts a decoupled detection head for efficient multi-scale learning, YOLOv9 enhances contextual modeling through refined convolutional and neck structures, and YOLOv11 further improves performance with lightweight attention and optimized connectivity, leading to superior accuracy in our experiments. Although YOLOv12 incorporates additional modular and normalization enhancements, it does not surpass YOLOv11 on this dataset but demonstrates stable inference behavior. Overall, the comparative analysis highlights how architectural optimization can improve detection performance without disproportionately increasing model complexity, which is critical for edge and field-deployable systems.

3.5. Evaluation Metrics

For sign-language object detection, we evaluate using Precision, Recall, F1-score, and mAP50–95. Predictions are matched with ground-truth boxes via IoU (Intersection over Union), which is defined as the overlap area divided by the union area. A prediction is considered a True Positive (TP) when the class is correct and the IoU meets the threshold (e.g., 0.50); duplicate detections or incorrect classes are False Positives (FP), and unmatched ground truths are False Negatives (FN). True Negatives (TN) are not used in standard detection evaluation. Precision = TP / (TP + FP) measures accuracy, Recall = TP / (TP + FN) measures completeness, and F1-score = 2·(Precision·Recall) / (Precision + Recall) balances these metrics.

For mAP50–95, we calculate Average Precision (AP) for each class as the area under the Precision–Recall curve, then average AP across IoU thresholds from 0.50 to 0.95 in steps of 0.05 (COCO-style), and across classes to obtain mAP. We report both mAP@50 (more lenient matching) and mAP50–95 (more strict, comprehensive), along with macro-averaged Precision, Recall, F1, and, ideally, standard deviations or confidence intervals to demonstrate the robustness of the results.

In addition to reporting global mAP scores, class-wise precision, recall, and AP values were computed to assess whether performance was biased toward high-frequency gesture categories. To account for result variability, each model was trained three times with different random seeds, and standard deviation across runs is reported for the main metrics. Full per-class AP values are provided in the supplementary material, while summary statistics for the top and bottom five classes are included in the Results section.

4. Results and Discussion

4.1. Model Performance

This section presents the initial performance results of YOLOv8, YOLOv9, YOLOv11, and YOLOv12, with a focus on accuracy, efficiency, and stability in real-world scenarios. Quantitative outcomes and illustrative examples appear in figures 4 through 5 and table 2, highlighting trends, strengths, sensitivities, and cross run consistency as the models contend with variable lighting, diverse viewpoints, partial occlusions, and motion.

After applying class-weighted loss during training, the model showed reduced prediction bias toward high-frequency classes. The improvement was reflected in higher recall values for low-frequency classes without degrading the performance of majority classes. While the results indicate that the mitigation strategy partially alleviated imbalance effects, further refinements such as per-class augmentation or focal loss may be required for deployment in real-world clinical settings where rare symptoms are equally important for diagnosis.

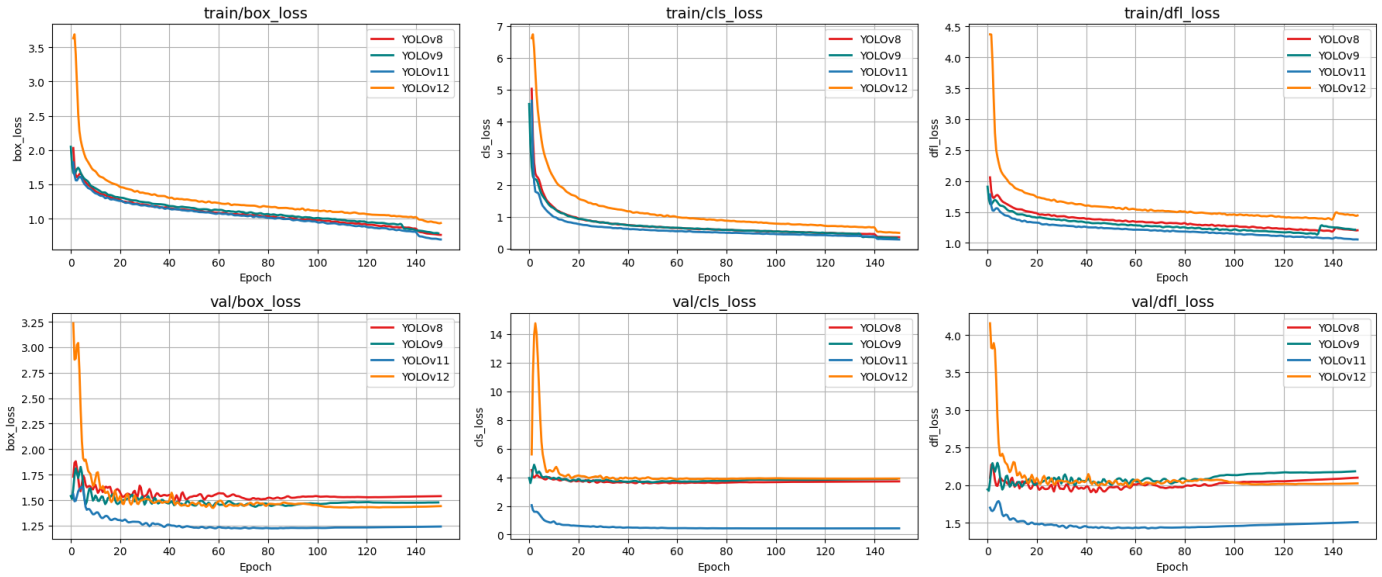

Figure 4. Training and Validation Loss Performance Comparison

 ISSN 2723-6471

The figure 4 depicting the training and validation loss trends illustrates the convergence behavior of the YOLOv8, YOLOv9, YOLOv11, and YOLOv12 models over 100 training epochs. Three loss components are considered: box loss for object localization accuracy, classification loss for class prediction error, and distribution focal loss for refining bounding box regression in anchor-free architectures. The training loss curves show that YOLOv9 and YOLOv8 achieve a faster and more consistent reduction in all loss components, indicating stable optimization and effective learning. In contrast, YOLOv11 and YOLOv12 demonstrate slower convergence and greater fluctuations, particularly in classification loss. Analysis of the validation losses further reveals that YOLOv9 maintains the lowest and most stable values across all metrics, reflecting strong generalization and limited overfitting. YOLOv8 exhibits comparable validation performance with slightly higher variability in distribution focal loss, while YOLOv11 and YOLOv12 display higher and less stable validation losses, which correspond to their lower detection accuracy. Overall, the results suggest that YOLOv9 offers the most effective error minimization, YOLOv8 provides a balanced trade-off between accuracy and efficiency, and the lighter architectures compromise precision for reduced complexity.

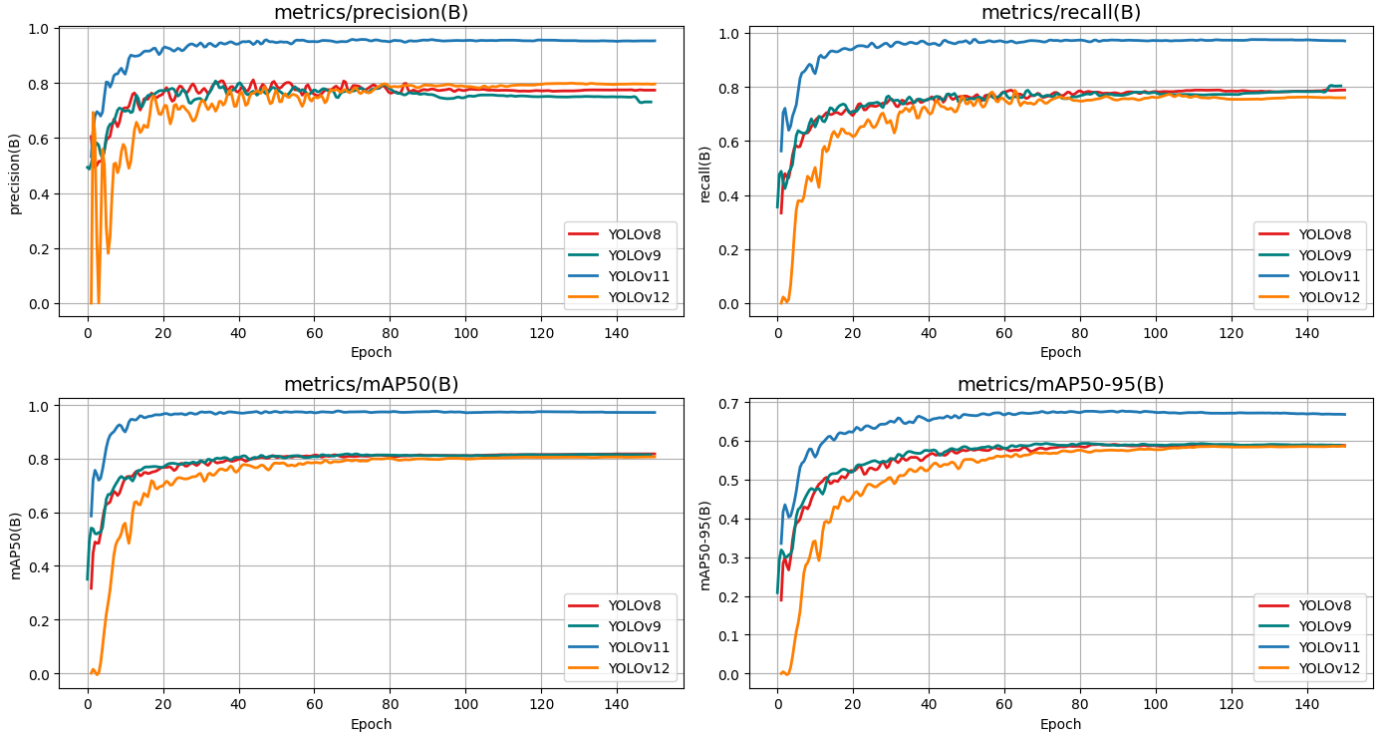

Figure 5. Comparative Analysis of Performance Metrics

Figure 5 the preliminary performance evaluation of YOLOv8, YOLOv9, YOLOv11, and YOLOv12, emphasizing their accuracy, computational efficiency, and robustness under practical conditions. YOLOv9 shows superior accuracy in most metrics, while YOLOv8 provides stable, balanced performance.

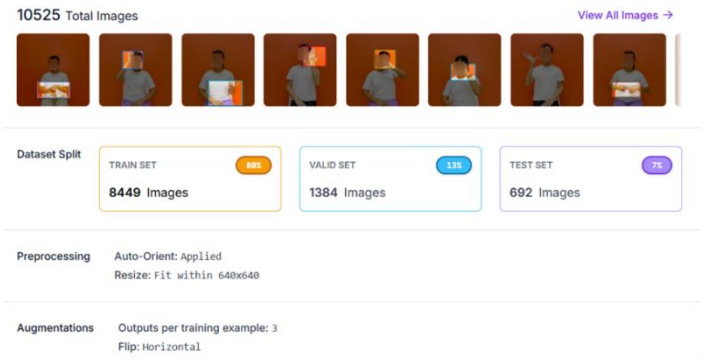

Figure 6. The image depicts the model in testing.

 ISSN 2723-6471

Figure 6 provides an overview of the dataset and evaluation setup, consisting of 10,525 images partitioned into training (8,449 images, 80%), validation (1,384 images, 13%), and testing (692 images, 7%) subsets. Preprocessing includes automatic orientation adjustment and resizing to 640×640 pixels, while horizontal flipping is applied as a data augmentation strategy to improve model robustness.

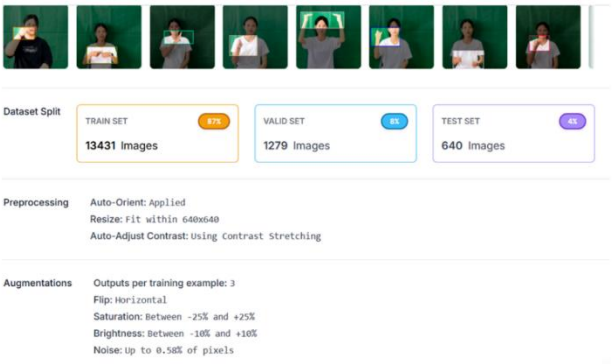

Figure 7. The image depicts the model in testing.

Figure 7 presents an overview of the dataset setup employed for model development and evaluation. A total of 15,350 images are allocated to the training set (13,431 images, 87%), validation set (1,279 images, 8%), and test set (640 images, 4%). Preprocessing steps include automatic orientation adjustment, resizing to 640×640 pixels, and contrast enhancement. In addition, data augmentation techniques such as horizontal flipping, brightness and saturation variation, and noise addition are applied to increase data variability and strengthen model robustness.

Table 2 provides a numerical comparison of four YOLO variants in terms of detection accuracy, model size, and inference efficiency, highlighting distinct performance trade-offs. YOLOv11 delivers the strongest detection results, achieving precision of 0.9527, recall of 0.9700, and mAP values of 0.9723 at IoU 0.5 and 0.6679 across 0.5–0.95. While its parameter count remains moderate at 7.7 million and model size is compact at 7.8 MB, its inference latency increases to 16.1 ms, reflecting higher computational requirements.

Table 2. Model Performance

Model	Precision	Recall	mAP@50	mAP@50:95	Params (M)	Size (MB)	Latency (ms)
YOLOv8	0.7735	0.7884	0.8178	0.5878	3.2	11.2	11.4
YOLOv9	0.7307	0.8037	0.8149	0.5883	24.2	12.2	13.7
YOLOv11	0.9527	0.9700	0.9723	0.6679	7.7	7.8	16.1
YOLOv12	0.7957	0.7598	0.8069	0.5856	20	22	20.5

YOLOv8 and YOLOv9 exhibit similar accuracy levels but differ markedly in efficiency. YOLOv8 attains precision and recall of 0.7735 and 0.7884, with an mAP@0.5–0.95 of 0.5878, while offering the lowest latency at 11.4 ms and the smallest parameter count of 3.2 million. YOLOv9 slightly improves recall to 0.8037 and mAP@0.5–0.95 to 0.5883 but requires a substantially larger model with 24.2 million parameters and higher latency of 13.7 ms. YOLOv12 records moderate performance with an mAP@0.5–0.95 of 0.5856 but incurs the highest latency at 20.5 ms and a larger model size of 22 MB, indicating limited efficiency gains from increased complexity. Overall, the results numerically confirm the trade-off between accuracy and efficiency, positioning YOLOv11 for accuracy-driven tasks and YOLOv8 as a practical choice for real-time or resource-limited applications.

4.1.1. Error and Failure Case Analysis

In addition to the overall performance metrics, failure-case analysis was conducted to identify the conditions under which the detector struggled. Most misclassifications occurred in samples with low illumination, motion blur, or partial hand occlusion, particularly when the dominant signing hand overlapped the torso or clothing. The lowest AP values were observed in semantically related categories such as Stress Level and Sleep Pattern, where gestures share similar

ISSN 2723-6471

hand shapes and differ mainly by spatial positioning. Figure 8 presents examples of failed detections, showing that the model occasionally confuses visually similar signs or fails to localize the hand when lighting causes loss of contour contrast. These findings suggest that incorporating temporal features or pose-estimation cues may reduce gesture ambiguity and improve robustness in real medical settings, where camera placement and environmental lighting cannot be strictly controlled.

Figure 8. Example of failed detections.

4.2. Comparative Discussion of Results and the Literature

While YOLO-based architectures were selected as the primary detection backbone in this study, alternative detection models were also reviewed to ensure that the choice was not constrained solely by accuracy benchmarks. EfficientDet, for example, offers strong accuracy–parameter scaling but exhibits higher inference latency on edge devices, limiting its suitability for real-time interaction in clinical settings. Transformer-based detectors such as DETR and Deformable DETR provide improved global feature reasoning but require longer convergence times, larger GPU memory, and typically underperform on small, high-detail objects such as hand shapes unless heavily optimized. Similarly, full transformer sequence models achieve strong performance in continuous sign recognition tasks, but their computational cost makes them impractical for deployment in standalone medical devices. Accordingly, YOLOv11 was identified as the most suitable model, as it delivered the highest detection performance in the experiments while maintaining a compact parameter size and stable validation loss. This combination indicates efficient learning and strong generalization, making YOLOv11 well suited for edge-oriented deployments such as mobile aquaculture tools or on-site biological monitoring systems, where limited hardware resources and low-latency inference are essential.

Previous studies consistently show that lightweight detectors such as YOLOv9-tiny and YOLOv11-nano can achieve near-optimal performance in resource-constrained settings [4], [9]. Although YOLOv8 exhibits slightly lower accuracy, it maintains a favorable balance between architectural simplicity and inference stability, supporting its use in general-purpose detection tasks, in line with earlier benchmarks on YOLO-based models for real-time, low-resource applications [4], [9]. Unlike multimodal approaches that incorporate skeletal features or sensor fusion [7], this work focuses exclusively on image-based classification to better capture visual characteristics relevant to species identification; however, known robustness issues under occlusion and varying illumination remain, particularly for juvenile specimens with subtle morphological differences [9]. Addressing a domain rarely explored in computer vision, this study targets the taxonomic complexity of economically important freshwater snails in Thailand and extends beyond prior mobile applications centered on generic gesture or object recognition [6], [8] by offering a scalable automated framework for aquaculture monitoring, traceability, and sustainable management

4.3. Connections and Future Applications

The steady improvements from YOLOv5 through YOLOv10 demonstrate that our detector is a reliable “perception core" for integrating with pose/keypoint streams and text modules, enabling comprehensive Thai Sign Language (TSL) communication. Results align with clinical workflows: the detector provides trustworthy visual evidence, pose cues help stabilize articulation during occlusion or movement, and downstream NLP/OCR translates signs into medical intentions, symptoms, and structured data for history-taking. This system supports triage kiosks, telemedicine, and low-resource clinics with limited interpreters. Short-term applications include registration, symptom intake, and captioning; mid-term objectives involve continuous sign recognition, context-aware intent linking with EHRs, and interactive clarification to minimize risks. For deployment, models should be compressed and distilled for on-device inference, data should be expanded to include dialectal and co-articulated TSL, and the system should be stress-tested in various

 ISSN 2723-6471

conditions and scenarios. Evaluation should go beyond mAP to include sequence-level accuracy and task outcomes such as intake completeness and error severity, with privacy-focused logging and governance to ensure safety.

5. Discussion

Experiments spanning YOLOv5 to YOLOv10 show steady progress, with v10 providing a credible balance between accuracy and computational limits. The YOLO architecture enhances Thai Sign Language recognition in challenging conditions, such as varying lighting, different viewpoints, partial occlusion, and rapid motion. However, gains in mAP at the 50 to 95 thresholds remain modest, indicating ongoing issues with IoU consistency and environmental variability. YOLOv11 was not included in the final analysis due to training and validation problems, despite promising early signs.

The findings point to a development path starting with a solid vision core and expanding into multimodal capabilities. The first step involves a reliable image recognition backbone. This robustness can be improved through pose estimation, keypoint detection, and integration with optical character recognition and natural language processing. This layered approach helps address occlusion, variation between individuals, and different viewpoints. Connecting recognition to clinical workflows further supports systematic history taking, standardized medical record keeping, and integration with health informatics systems.

Implementing real-world deployment requires thoughtful model compression, including pruning, quantization, and knowledge distillation, to reduce model size and latency on resource-limited edge devices. This is crucial for rural clinics and service points where power and connectivity are inconsistent.

This study has some limitations. Dataset diversity remains limited; protective gear and hand coverings can hide important cues, and the evaluation still focuses on single frames instead of sequences. Future research should include sequence-level metrics such as sequence mAP and sequence F1, conduct field tests in real environments, measure device latency, and enforce thorough data governance, including privacy safeguards, informed consent, de-identification, and secure data storage. Overall, the practical approach is to build a reliable vision core, extend it to multimodal processing, and adapt it for edge deployment, paving the way for sustainable clinical readiness in Thai settings.

Ethical safeguards are essential when deploying an AI-based sign language system in clinical environments. All video data used in this study were collected under informed consent agreements approved by the institutional ethics committee, and signer identities were anonymized through face-blurring and metadata removal to comply with the Thai Personal Data Protection Act (PDPA). Dataset access is restricted through encrypted storage and role-based authorization. Because gesture misclassification may lead to incorrect symptom reporting and clinical misunderstanding, the system is designed as an assistive tool rather than an autonomous diagnostic agent; all generated outputs are reviewed by medical personnel before being entered into patient records. Future deployment will require real-time explainability logs, audit trails, and a fallback mechanism that allows patients to request human interpretation when model confidence is low, reducing the risk of medical harm in high-stakes interactions.

6. Conclusion

This experimental evaluation reveals clear performance variations among YOLOv8, YOLOv9, YOLOv11, and YOLOv12 under identical training conditions. As shown in Table 2, YOLOv11 consistently outperforms the other models, achieving the highest Precision, Recall, mAP@50, and mAP@50:95. While its inference latency is higher than that of YOLOv8 and YOLOv9, YOLOv11 maintains a relatively compact model size and moderate parameter count, indicating a favorable trade-off between accuracy and computational efficiency. In comparison, YOLOv8 and YOLOv9 exhibit lower detection accuracy, and YOLOv12 shows reduced effectiveness alongside increased latency and model size, suggesting limited benefits from additional architectural complexity.

These results support the selection of YOLOv11 as the backbone of the proposed framework, particularly for applications where detection accuracy and robustness are critical. The findings further suggest a modular development approach in which a reliable image-based detection model forms the foundation, with future integration of pose estimation or keypoint-based features to enhance performance under real-world conditions such as occlusion, motion,

illumination variation, and signer diversity. Moreover, combining accurate detection outputs with higher-level semantic interpretation enables practical downstream applications, including automated analysis and structured system integration in assistive and real-world deployment settings.

Future work will focus on extending the framework beyond isolated sign detection toward sequence-level recognition, expanding the dataset to include dialectal variants, co-articulated signing, and richer contextual cues, and applying model compression and optimization techniques to enable reliable on-device inference in resource-limited clinical environments. These developments are essential for ensuring deployability, low latency, and clinical safety in real Thai healthcare settings.

7. Declarations

7.1. Author Contributions

Conceptualization: P.N., P.K., P.J., K.S.N., K.S.N., and W.S.N.; Methodology: K.S.N.; Software: P.N.; Validation: P.N., K.S.N., and W.S.N.; Formal Analysis: P.N., K.S.N., and W.S.N.; Investigation: P.N.; Resources: K.S.N.; Data Curation: K.S.N.; Writing Original Draft Preparation: P.N., K.S.N., and W.S.N.; Writing Review and Editing: K.S.N., P.N., and W.S.N.; Visualization: P.N.; All authors have read and agreed to the published version of the manuscript.

7.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

7.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

7.4. Institutional Review Board Statement

Not applicable.

7.5. Informed Consent Statement

Not applicable.

7.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] K. Xia, W. Lu, H. Fan, and Q. Zhao, "A Sign Language Recognition System Applied to Deaf-Mute Medical Consultation," *Sensors,* vol. 22, no. 23, pp. 1-17, Jan. 2022, doi: 10.3390/s22239107

[2] C. J. Moreland, S. R. Rao, K. Jacobs, and P. Kushalnagar, "Equitable Access to Telehealth and Other Services for Deaf People During the COVID-19 Pandemic," *Health Equity,* vol. 7, no. 1, pp. 126–136, 2023, doi: 10.1089/heq.2022.0115.

[3] I. Papastratis, C. Chatzikonstantinou, D. Konstantinidis, K. Dimitropoulos, and P. Daras, "Artificial Intelligence Technologies for Sign Language," *Sensors (Basel)*, vol. 21, no. 17, pp. 1-12, , Aug. 2021, doi: 10.3390/s21175843.

[4] B. Alsharif, E. Alalwany, A. Ibrahim, I. Mahgoub, and M. Ilyas, "Real-Time American Sign Language Interpretation Using Deep Learning and Keypoint Tracking," *Sensors,* vol. 25, no. 7, pp. 21-38, Jan. 2025, doi: 10.3390/s25072138.

[5] J. Huang and V. Chouvatut, "Video-Based Sign Language Recognition via ResNet and LSTM Network," *Journal of Imaging,* vol. 10, no. 6, pp. 1-14, June 2024, doi: 10.3390/jimaging10060149.

[6] Y. Zhang, Y. Han, Z. Zhu, X. Jiang, and Y. Zhang, "Artificial intelligence in sign language recognition: A comprehensive bibliometric and visual analysis," *Computers and Electrical Engineering,* vol. 120, no. Dec., pp. 1-12, , Dec. 2024, doi: 10.1016/j.compeleceng.2024.109854.

[7] C. Lu, M. Kozakai, and L. Jing, "Sign Language Recognition with Multimodal Sensors and Deep Learning Methods," *Electronics,* vol. 12, no. 23, pp. 1-27, Jan. 2023, doi: 10.3390/electronics12234827.

[8] L. R. Soenksen, "Integrated multimodal artificial intelligence framework for healthcare applications," *npj Digit. Med.,* vol. 5, no. 1, pp. 149-164, Sept. 2022, doi: 10.1038/s41746-022-00689-4.

[9] M. K. Siam, M. J. Hossain Faruk, B. He, J. Q. Cheng, and H. Gu, "Multimodal Models in Healthcare: Methods, Challenges, and Future Directions for Enhanced Clinical Decision Support," *Information,* vol. 16, no. 11, pp. 971-988, Nov. 2025, doi: 10.3390/info16110971.

[10] N. Ardic and R. Dinc, "Emerging trends in multi-modal artificial intelligence for clinical decision support: A narrative review," *Health Informatics J,* vol. 31, no. 3, pp. 1-21, July 2025, doi: 10.1177/14604582251366141.

[11] W. S. Nuankaew, N. Nuttaphum, T. Sararat, P. Banyaem, and P. Nuankaew, "Sign Language Detection Mobile Application for Thai Patients Using Medical Image Processing to Support Medical Consultations," in Smart Innov. Syst. Technol., In C.S., Londhe N.S., Bhatt N., and Kitsing M., Eds., Springer Science and Business Media Deutschland GmbH, 2025, pp. 105–117. doi: 10.1007/978-981-96-1206-2_10.

[12] W. Vijitkunsawat, T. Racharak, C. Nguyen, and N. L. Minh, "Video-Based Sign Language Digit Recognition for the Thai Language: A New Dataset and Method Comparisons," *in ICPRAM 2023 - Proceedings of the 12th International Conference on Pattern Recognition Applications and Methods, Volume 1, Science and Technology Publications, Lda, 2023*, vol. 1, no. 1, pp. 775–782, 2023. doi: 10.5220/0011643700003411.

[13] M. Al-Qurishi, T. Khalid, and R. Souissi, "Deep Learning for Sign Language Recognition: Current Techniques, Benchmarks, and Open Issues," *IEEE Access*, vol. 9, no. 1, pp. 126917–126951, 2021, doi: 10.1109/ACCESS.2021.3110912.

[14] M. Aly and I. S. Fathi, "Recognizing American Sign Language gestures efficiently and accurately using a hybrid transformer model," *Sci Rep,* vol. 15, no. 1, pp. 1-13, June 2025, doi: 10.1038/s41598-025-06344-8.

[15] F. Roelofsen, L. Esselink, S. Mende-Gillings, and A. Smeijers, "Sign Language Translation in a Healthcare Setting," in Proceedings of the Translation and Interpreting Technology Online Conference, R. Mitkov, V. Sosoni, J. C. Giguère, E. Murgolo, and E. Deysel, Eds., Held Online: INCOMA Ltd., July 2021, pp. 110–124. Accessed: Aug. 25, 2025. [Online]. Available: https://aclanthology.org/2021.triton-1.13/

[16] F. M. Najib, "Sign language interpretation using machine learning and artificial intelligence," *Neural Comput & Applic,* vol. 37, no. 2, pp. 841–857, Jan. 2025, doi: 10.1007/s00521-024-10395-9.

[17] W. S. Nuankaew, T. Sararat, N. Lankham, A. Doksak, and P. Nuankaew, "Mobile Application for Tracking and Assisting the Visually Impaired Students' University of Phayao," *in Lect. Notes Networks Syst., Iglesias A., Shin J., Patel B., and Joshi A., Eds., Springer Science and Business Media Deutschland GmbH,* vol. 2025, no. 1, pp. 273–283, 2025. doi: 10.1007/978-981-96-1741-8_24.

[18] K. V. Reddy, G. Saketh, S. S. Priyanshu, M. Nitesh, S. K. Hussain, and K. V. Sharma, "AI-Driven Healthcare Management Platform: Enhancing Accessibility, Efficiency, and Security in Digital Health Systems," *Synthesis: A Multidisciplinary Research Journal,* vol. 3, no. 1, pp. 1–14, Mar. 2025, doi: 10.70162/smrj/2025/v3/i1/v3i101.

[19] A. Frosolini, L. Franz, V. Caragli, E. Genovese, C. de Filippis, and G. Marioni, "Artificial Intelligence in Audiology: A Scoping Review of Current Applications and Future Directions," *Sensors (Basel),* vol. 24, no. 22, pp. 1-26, Nov. 2024, doi: 10.3390/s24227126.

[20] J. N. Mandrekar, "Measures of Interrater Agreement," *Journal of Thoracic Oncology,* vol. 6, no. 1, pp. 6–7, Jan. 2011, doi: 10.1097/JTO.0b013e318200f983.

[21] M. Li, Q. Gao, and T. Yu, "Kappa statistic considerations in evaluating inter-rater reliability between two raters: which, when and context matters," *BMC Cancer,* vol. 23, no. 1, pp. 799-817, Aug. 2023, doi: 10.1186/s12885-023-11325-z.

[22] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *J Big Data,* vol. 6, no. 1, pp. 60-72, July 2019, doi: 10.1186/s40537-019-0197-0.

[23] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: Fast and Flexible Image Augmentations," *Information,* vol. 11, no. 2, pp. 125-137, Feb. 2020, doi: 10.3390/info11020125.

[24] H. Harrouch, L. Trabelsi, M. Jebali, and O. Gammoudi, "A Deep Learning-Based Method Combines Manual and Non-Manual Features for Sign Language Recognition," *Sci Rep,* vol. 2025, no. Dec., pp. 1-12, Dec. 2025, doi: 10.1038/s41598-025-32768-3.

[25] W. S. Nuankaew, Y. Yangwattana, S. Kamwaree, T. Sararat, and P. Nuankaew, "Harnessing AI for Agriculture: Plant Pest Detection on Web Application Using Deep Learning," *in Int. Conf. Digit. Arts, Media Technol., DAMT ECTI North. Sect. Conf. Electr., Electron., Comput. Telecommun. Eng., NCON, Institute of Electrical and Electronics Engineers Inc.*, vol. 2025, no. 1, pp. 588–593, 2025. doi: 10.1109/ECTIDAMTNCON64748.2025.10961964.